Sommaire

Table des figures

Liste des tableaux

Production scientifique

Papiers acceptés

• **Schramm C**, Vial C, Bachoud-Lévi A-C, Katsahian S. CLEB : a new method for treatment efficacy clustering in longitudinal data. *Statistical Methods in Medical Research.* 2015 Dec 31. (**Chapitre 3**)

• **Schramm C**, Katsahian S, Youssov K., Demonet J-F., Krystkoviak P., Supiot F., Verny C., Cleret de Langavant L., EHDI & MIGHD study groups, Bachoud-Levi AC. How to capitalize on the retest effect in future trials on Huntington's disease. *PLoS One.* 2015 Dec 29 ;10(12) :e0145842. (**Chapitre 7**)

Papiers soumis

• Diego-Balaguer R*, **Schramm C***, Rebeix I, Dupoux E, Dürr A, Brice A, Cleret de Langavant L, Youssov K, Fenelon G, Verny C, Demotte V, Azulay J.P, Goizet C, Kryskowiak P, Tranchant C, Maison P, Rialland A, Schmitz D, Fenelon G, Jacquemot C, French Speaking Huntington Group, Fontaine B, Bachoud-Lévi A.C. Predicting Huntington's disease progression : the influence of the COMT Val^{158}Met polymorphisms in a longitudinal prospective study. *Soumis à PLoS One, révision mineure.* (**Chapitre 5**)
*co-premiers auteurs

Papiers en perspective

• **Schramm C**, Jannot AS, Nevoret C, Katsahian S. Guidelines for using marker-based desings. *Simulations en cours.* (**Chapitre 6**)

• **Schramm C**, Diao G, Katsahian S. Clustering for treatment efficacy in recurrent events data. *En cours d'écriture, sera soumis à Statistical Methods in Medical Research* (**Annexe B**)

• Bachoud-Levi AC, ..., **Schramm C**, ... Fetal Striatal allograft in Huntington's disease : a multicentre, randomized, delayed start, phase 2 open-label trial . *En cours d'écriture, sera soumis à Lancet Neurology.*

• **Schramm C**, Varet H, Diao G, Vial C, Katsahian S. LongiClust : a R package for treatment efficacy clustering in longitudinal and recurrent events data. *Package programmé, en phase de test.*

Conférences internationales

Présentations orales

• ISCB 2014 (35th Annual Conference of the International Society for Clinical Biostatistics), Vienne, Autriche, 24-28 Août 2013, Clustering for treatment effect on recurrent events.

• SFDS, Biopharma 2013 (7th International Meeting Statistical Methods in Biopharmacy), Paris, France, 15-16 Septembre 2013, Clustering for treatment effect in longitudinal study : a new method for personalized medicine.

• ISCB 2013 (34th Annual Conference of the International Society for Clinical Biostatistics), Munich, Allemagne, 25-29 Août 2013, Clustering for treatment effect in longitudinal study : a new method for personalized medicine.

Posters

• Neurostemcell 2013, Bellagio, Italie, 13-15 Avril 2013, Clustering for graft effect in MIG-HD trial : a new statistical method.

Liste des abréviations

- AG : Andersen and Gill model
- CAG : Cytosine-Adénine-Guanine
- CAH : Clustering Ascendant Hiérarchique
- COMT : Cathecol-O-MethylTransférase
- CLEB : Clustering for Longitudinal data using Extended baseline
- EM : Espérance-Maximisation
- FAS : Functional Assessment Scale
- GT-UR : Gap Time - UnRestriceted model
- HVLT : Hopkins Verbal Learning Test
- IRM : Imagerie par Résonance Magnétique
- IS : Independance Scale
- IT15 : Interesting Transcript 15
- KML : K-Means for Longitudinal data
- LCMM : Latent-Class Mixed Model
- *log* : fonction logarithme népérien
- M- : Marqueur prédictif et/ou pronostique négatif

- M+ : Marqueur prédictif et/ou pronostique positif
- MADRS : Montgomery and Asberg Depression Rating Scale
- MDRS : Mattis Dementia Rating Scale
- Met : Méthionine
- ML : Maximum Likelihood
- MMSE : Mini Mental Status Examination
- MIGHD : Multicentric Intracerebral Graft in Huntington's Disease
- $\mathcal{N}(\mu,\sigma^2)$: loi normale de moyenne μ est de variance σ^2
- REML : REstricted Maximum Likelihood
- RHLF : Réseau Huntington de Langue Française
- SDMT : Symbol Digit Modalities Test
- TFC : Total Functional Capacity
- TMS : Total Motor Score
- TMT : Trail Making Test
- UHDRS : Unified Huntington's Disease Rating Scale
- Val : Valine

Introduction : Les problèmes statistiques posés par les études de biothérapie dans la maladie de Huntington

Les données longitudinales sont des mesures d'une même variable, chez les mêmes patients, au cours du temps. Le traitement statistique des données longitudinales doit tenir compte de la variabilité intra-patients et la variabilité inter-patients comme sources d'hétérogénéité des données. Lorsque les données proviennent d'études observationnelles ou d'essais cliniques, elles permettent de mettre en évidence des marqueurs pronostiques de l'évolution de la maladie et des marqueurs prédictifs de l'efficacité du traitement contre la progression de la maladie. La validation de ces marqueurs nécessite des méthodes statistiques pour (i) identifier des sous-groupes de patients et (ii) concevoir des essais cliniques adaptés.

Nous nous sommes intéressés à ces questions dans le cadre spécifique des petits effectifs, avec comme application les biothérapies dans la maladie de Huntington. Cette maladie rare est multifacette et de durée d'évolution longue, induisant une grande hétérogénéité entre les patients que ce soit sur la présentation de la maladie ou sur son évolution. Les biothérapies en cours d'essai pour cette maladie sont réalisées sur des petits effectifs, avec un effet mesurable à long terme et hétérogène. Identifier des marqueurs d'évolution de la maladie et de réponse au traitement permettrait de mieux comprendre et d'améliorer les résultats des études de biothérapie dans la maladie de Huntington.

La maladie de Huntington

La maladie de Huntington est une maladie neurodégénérative génétique rare orpheline et se traduit cliniquement par des troubles moteurs (mouvements anormaux involontaires, trouble de l'équilibre,...), cognitifs (perte de mémoire, désorientation dans l'espace,...), et/ou psychiatriques (dépression, irritabilité,...). La maladie se déclare en moyenne autour

de 30-50 ans et les troubles s'accumulent progressivement entraînant une perte d'autonomie et conduisant au décès du patient en 15 à 20 ans. On définit cinq stades de la maladie [1] :

- Stade 1 : vie familiale et professionnelle normale, parfois des problèmes comportementaux
- Stade 2 : possible vie professionnelle avec facultés réduites, accomplissement des tâches de la vie quotidienne avec quelques difficultés, apparition des premiers symptômes graves
- Stade 3 : impossibilité de travailler, de faire des tâches ménagères et de gérer des affaires financières courantes, altération des fonctions vitales
- Stade 4 : impossibilité d'accomplir seul les activités de la vie quotidienne, aide professionnelle minimale, communication verbale impossible
- Stade 5 : besoin d'une aide permanente pour toutes les activités de la vie quotidienne, nécessité de séjour dans un centre de soins prolongés, communication pratiquement nulle

La maladie de Huntington est une maladie génétique autosomique dominante due à la mutation du gène IT15 sur le bras court du chromosome 4 (4p16.3), codant la Huntingtine (Htt). Ce gène contient de 6 à 35 répétitions du trinucléotide Cytosine-Adénine-Guanine (CAG) et le nombre de répétitions est augmenté dans le cas de la maladie de Huntington [2]. La pénétrance de la maladie varie en fonction du nombre de répétitions de CAG [3]. La pénétrance est incomplète de 36 à 40 répétitions et complète à partir de 40 répétitions, c'est-à-dire que tous les individus exprimeront le phénotype de la maladie au cours de leur vie [4]. De 27 à 35 répétitions, on parle de cas intermédiaires, car les sujets, bien qu'ils ne manifestent pas les signes de la maladie, pourraient transmettre la mutation à leurs enfants [5]. Plus le nombre de répétitions est important, plus la maladie apparaîtra précocement et plus sa progression sera rapide [6, 7]. Lorsque la maladie se développe avant 20 ans (souvent associée à plus de 60 répétitions), on parle de forme juvénile [8, 9].

Il est possible de réaliser un test génétique afin de savoir si l'on est porteur de la maladie. Cette demande est encadrée par une équipe pluridisciplinaire (généticien, psychiatre, neurologue) et se déroule sur plusieurs mois du fait de l'impact du résultat sur le sujet à risque et sa famille [10].

La physiopathologie de la maladie reste inconnue à ce jour, mais les recherches ont montré un rôle protecteur de la protéine Htt et un rôle délétère de la protéine huntingtine mutée (Httm) dans le cerveau. La Htt interviendrait dans le transport de vésicules contenant un facteur neurotrophique essentiel à la survie des neurones [11]. Dans le cas de la maladie de Huntington, la Htt formerait des agrégats entravant les fonctions normales de la protéine et induisant la mort neuronale. Les régions les plus atteintes sont les ganglions de la base (notamment le striatum, voir Figure 1), puis le cortex (couches périphériques

du cerveau) au fur et à mesure de l'évolution de la maladie. A terme, on observe une atrophie dans toutes les structures du cerveau.

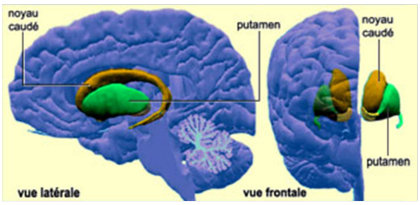


FIGURE 1 – Le striatum

Le striatum, au centre du cerveau, est composé du noyau caudé et du putamen.

Un suivi longitudinal de la maladie grâce au centre de référence et aux centres de compétences pour les maladies rares

La prévalence de la maladie de Huntington est de 2,71 [1,55 - 4,72] malades pour 100 000 au sein de la population mondiale, mais varie géographiquement, de 0,25 [0,14 - 0,42] pour la Chine à 12,08 [9,08 - 15,76] pour l'Australie [12]. En France, on estime le nombre de personnes atteintes de la maladie de Huntington à environ 6000 (soit 9 personnes pour 100 000) tandis que 12 000 personnes pourraient être porteuses du gène muté.

La rareté et la complexité de la maladie sont des freins pour la recherche de nouveaux traitements mais aussi pour la prise en charge des patients. Cette maladie nécessite un suivi par des experts. En France, la mise en place des centres de référence et de compétence pour les maladies rares a permis de simplifier et d'intensifier les recherches sur ces maladies. Le centre national de référence pour la maladie de Huntington est situé sur quatre hôpitaux : l'hôpital Henri Mondor de Créteil pour le suivi des patients et la coordination, l'hôpital Albert Chenevier de Créteil pour les formes avancées, l'hôpital Armand Trousseau de Paris pour les enfants et l'Institut du Cerveau et de la Moelle épinière de Paris pour la génétique. Ces centres développent chacun des compétences spécifiques dans la maladie de Huntington et exercent une attraction interrégionale, nationale ou internationale, permettant le suivi d'une grande cohorte de patients. Dans cette maladie, le suivi longitudinal des patients est extrêmement important pour mieux appréhender leur déclin tout au long du processus de la maladie.

En France, la première cohorte de patients a débuté en 2002 avec le Réseau Huntington de Langue Française (RHLF), coordonné à l'hôpital Henri Mondor de Créteil, et comportant aujourd'hui plus d'un millier de patients. Cette cohorte a été intégrée à REGISTRY, la cohorte de l'*European Huntington's Disease Network* en 2005 qui elle-

même sera une composante de ENROLL, une cohorte mondiale, dès 2015. Au centre de référence Henri Mondor, nous avons accès aux données des patients francophones. Ces données regroupent les caractéristiques socio-démographiques et génétiques des patients ainsi que leurs antécédents personnels et familiaux. Les patients ont une visite annuelle où sont évaluées leurs capacités motrices, fonctionnelles et cognitives ainsi que leur état psychiatrique grâce à des échelles d'évaluation standardisées tel que l'UHDRS (*Unified Huntington's Disease Rating Scale*) [13]. Les échelles d'évaluation utilisées dans la maladie de Huntington sont détaillées en Annexe A.1. Ces échelles constituent des marqueurs de l'évolution de la maladie.

Les biothérapies

Actuellement, des traitements symptomatiques peuvent améliorer l'état des patients. Par exemple les neuroleptiques permettent de limiter les mouvements anormaux et les troubles psychiatriques tandis que les antidépresseurs peuvent prémunir les patients contre la dépression ou l'anxiété, d'autant que le risque de suicide est accentué par la maladie [14]. Cependant aucun traitement curatif n'existe. Les recherches de ces dernières années se tournent, entre autres, vers la neuroprotection et les biothérapies.

Les biothérapies sont une nouvelle classe de thérapeutiques regroupant à la fois les thérapies géniques (transfert de gènes, intervention sur les gènes) [15], les thérapies cellulaires ou tissulaires substitutives (manipulation de cellules souches ou différenciées) [16, 17, 18], et de manière générale tous les traitements modifiant les paramètres biologiques du patient. Cette classe thérapeutique bouleverse le paysage des essais cliniques. Trois aspects compliquent l'évaluation de l'efficacité du traitement. Premièrement, la complexité et les coûts engendrés par ces traitements impliquent de réaliser des essais cliniques sur de petits effectifs de patients. Deuxièmement, la multitude des étapes nécessaires à la mise en place du traitement, très dépendantes du patient, ajoute de la variabilité. Enfin, ces thérapies nécessitent des actes de chirurgie, rendant l'essai difficilement réalisable en aveugle. Bien que ce type d'essai en double aveugle ait déjà été utilisé [19], cela pose des problèmes d'éthique. En effet, un des critères du traitement de référence ou du placebo est qu'il ne doit pas nuire aux patients, les actes d'anesthésie et de chirurgie comprenant tous les deux des risques [20].

De plus, l'effet de ces traitements peut être lié aux caractéristiques individuelles du patient, qu'elles soient cliniques, génétiques, biologiques ou immunitaires, incitant à développer différentes stratégies thérapeutiques en parallèle et à définir pour chaque patient, celle qui lui sera favorable. Cela passe par une modification des plans expérimentaux utilisés dans les essais cliniques. On ne valide plus seulement le traitement mais aussi des marqueurs d'efficacité du traitement et l'utilité de recourir à une médecine stratifiée [21].

Les greffes de neurones pour la maladie de Huntington

L'une des biothérapies proposées dans plusieurs essais cliniques de phase I ou II sur la maladie de Huntington est la transplantation de cellule fœtales. L'allogreffe consiste à implanter dans le striatum atrophié des patients, des neurones homologues issus de l'éminence ganglionnaire, zone de formation du striatum, provenant de fœtus humain après interruption volontaire de grossesse. L'idée sous-jacente est que les neurones fœtaux se différencient en neurones striataux et établissent des connexions neuronales fonctionnelles. Les études sur les animaux ont montré un bénéfice de cette technique chez des macaques pour lesquels une lésion striatale a été induite [22] ainsi que chez des souris transgéniques exprimant le génotype de la maladie humaine [23]. Depuis 1998, sept études utilisant ce procédé chez les humains ont été publiées, incluant de 2 à 10 patients [24, 25, 26, 27, 28, 29, 30]. Les résultats restent variables aussi bien entre les études qu'au sein d'une même étude [18]. Les techniques utilisées diffèrent et aucune recommandation officielle n'existe. La première étude montrant un bénéfice chez des patients a été réalisée en France [25] en 2000 avec trois patients sur cinq ayant un bénéfice de la greffe à long-terme [31]. Les bonnes performances cliniques corrèlent avec le métabolisme striatal observé en fluorodesoxyglucose (FDG) par tomoscintigraphie par émission de positrons (TEP).

Les résultats encourageants de l'étude pilote réalisée à l'hôpital Henri Mondor de Créteil sur cinq patients [25, 31], ont conduit à réaliser un essai clinique contrôlé et randomisé chez un plus grand nombre de patients afin de démontrer l'efficacité de l'allogreffe. L'essai clinique multicentrique de greffe intracérébrale de neurones fœtaux pour le traitement de la maladie de Huntington (*Multicentric Intracerebral Grafting in Huntington's Disease*, MIG-HD) a commencé en 2001 (NCT00190450). C'est un essai ouvert en « *delayed-start* », comprenant trois phases pour une durée de suivi de 52 mois (Figure 2). En premier lieu, tous les patients sont suivis sans traitement pendant un an. Puis, les patients sont randomisés soit dans le groupe « greffe précoce » soit dans le groupe « greffe tardive », le groupe « greffe précoce », greffé à M13 et M14 et le groupe « greffe tardive », greffé à M33 et M34. Entre M12 et M32, le groupe « greffe tardive » constitue un groupe contrôle. Compte tenu de la durée de l'essai et du déclin pressenti des patients non traités, le comité d'éthique a jugé nécessaire que tous les patients soient greffés. Dans cet essai, le groupe contrôle n'a pas subi d'intervention placebo.

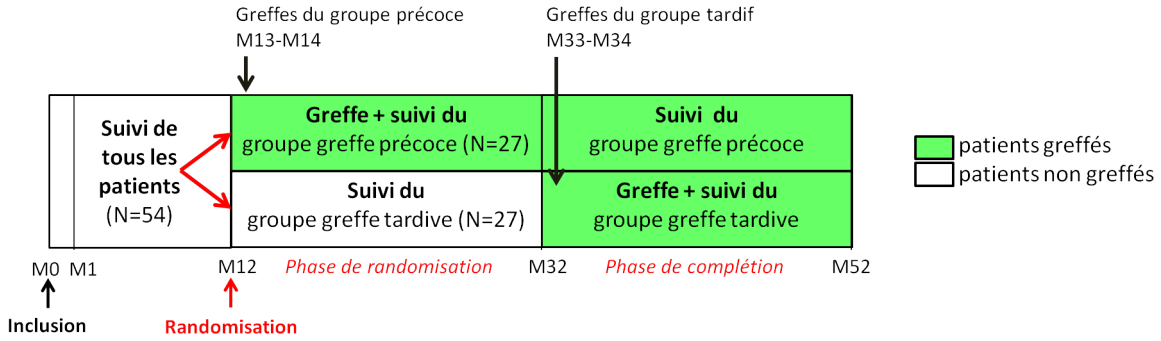


FIGURE 2 – Le plan expérimental de l'essai MIG-HD

Les patients sont inclus et testés pour la première fois à M0. Ils sont testés une seconde fois à M1 afin de limiter un éventuel effet d'apprentissage des tests. Des critères d'exclusion sont proposés à M1 et au moment de la randomisation (M12). Les patients ont un bilan complet à M0, M1, M12, M32 et M52 et des bilans intermédiaires environ tous les six mois, ainsi qu'un entretien avec le neurologue tous les trois mois. A chaque bilan complet, les données recueillies pour chaque patient comportent la clinique, l'imagerie, la biologie et l'électrophysiologie. Par patient et par bilan complet, le nombre de données, toutes sources confondues, s'élève à plus de 400 données brutes ce qui revient à plus de 200 000 données brutes pour les 54 patients inclus dans l'étude.

Problématiques statistiques liées à l'étude de l'efficacité des greffes

Hétérogénéité de l'efficacité du traitement : identification des répondeurs

La première question à l'origine de notre travail est la définition de patients répondeurs à la greffe. En effet, l'hétérogénéité intra-étude observée dans les précédents essais [18] semble se confirmer avec l'essai MIG-HD. Nous voulons développer une méthode d'apprentissage non supervisée (clustering) permettant de définir des sous-groupes (« clusters ») de patients, cette approche étant ensuite appliquée aux données de l'essai MIG-HD. Les méthode de clustering permettent d'identifier des sous-groupes de patients sans *a priori*. Le clustering cherche à maximiser la cohésion interne (c'est-à-dire minimiser la variabilité au sein de chaque cluster) et l'isolation externe (c'est-à-dire maximiser la variabilité entre les clusters) tel que le représente la Figure 3 [32].

Les premiers algorithmes de clustering, développés dès les années 1960, permettent de classer les patients selon des critères fixes. Nous nous intéressons plutôt à la progression naturelle de la maladie et la modification de son évolution grâce à un traitement. Nous

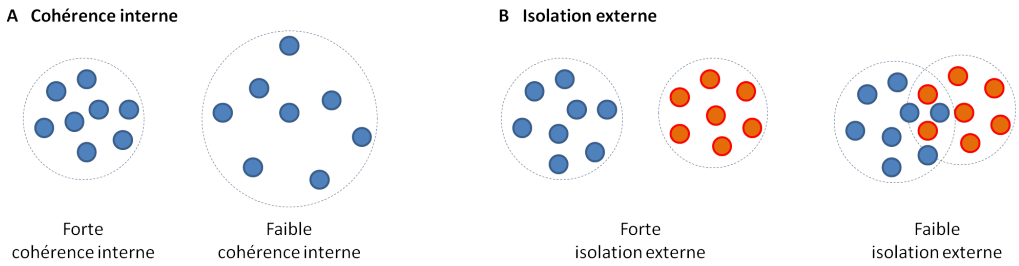


FIGURE 3 – Représentation schématique de l'isolation externe et de la cohérence interne

La cohérence interne mesure la ressemblance entre les individus d'un même groupe. L'isolation externe mesure la dissimilarité entre les individus de groupes différents. Un point bleu ou rouge représente un patient.

voulons garder le maximum d'information possible, et donc étudier l'entièreté de l'évolution des scores obtenus par le patient sur une période donnée. Cela nécessite l'utilisation de méthodes statistiques pour l'analyse des données longitudinales, car nous utilisons plusieurs mesures à des temps différents par patient. Des méthodes de clustering pour données longitudinales, paramétriques et non paramétriques, ont été développées ces quinze dernières années. On peut notamment citer les algorithmes des K-moyennes pour les données longitudinales [33, 34] et les modèles mixtes par classes latentes [35]. Ces méthodes sont de plus en plus utilisées dans le domaine biomédical [36, 37, 38, 39, 40]. Dans le cas de l'étude du bénéfice des greffes chez les patients Huntington, et dans toutes les études où l'effet du traitement est modélisé par un changement de pente, ces méthodes ne sont pas satisfaisantes car elle ne prennent pas en compte l'information de la pente pré-traitement qui est essentielle dans le cas où les patients n'ont pas le même profil d'évolution avant traitement. De plus, l'hétérogénéité intra-patient et l'hétérogénéité du déclin naturel s'ajoutent à l'hétérogénéité de l'effet du traitement.Ces deux sources additionnelles d'hétérogénéité sont à l'origine des profils d'évolution observés (Figure 4).

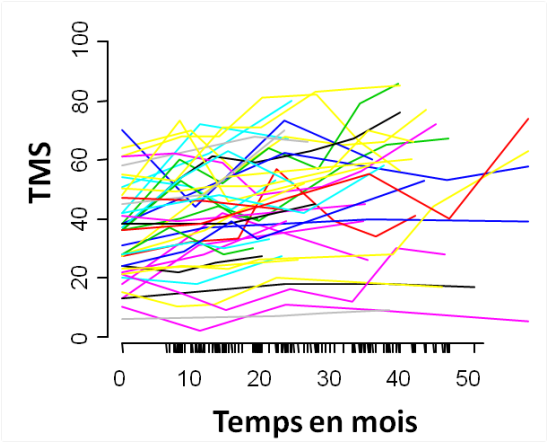


FIGURE 4 – Evolution du score moteur de l'UHDRS (*Total Motor Score*, TMS) chez des patients atteints de la maladie de Huntington

Ceci a motivé le développement d'une nouvelle méthode de clustering permettant d'identifier des sous-groupes de patients selon la réponse à un traitement quand l'effet de ce dernier se traduit par un changement dans la progression de la maladie. En particulier, elle doit pouvoir s'utiliser dans la recherche de patients répondeurs à la greffe dans le cas de la maladie de Huntington, c'est-à-dire tenir compte de toutes les sources d'hétérogénéité, y compris les délais entre deux visites et les données manquantes, et cela même pour de petits effectifs de patients. La notion de petits effectifs est importante pour deux raisons. La première est que la méthode doit pouvoir s'appliquer sur un petit échantillon de patient. La seconde est que la méthode doit pouvoir identifier un faible taux de répondeurs au sein d'un échantillon de patients.

Amélioration des plans expérimentaux des futurs essais cliniques

La seconde question à l'origine de notre travail est l'amélioration des plans expérimentaux pour les futurs essais cliniques de greffe dans la maladie de Huntington. Il s'agit notamment d'inclure si possible les marqueur prédictifs d'efficacité de la greffe, les marqueurs pronostiques de l'évolution de la maladie et les mesures du déclin cognitif. En effet, le critère de jugement principal pour la maladie de Huntington et pour les maladies neurodégénératives en général, repose, plus souvent, sur un score moteur ou fonctionnel (Annexe A.2, Table 10), que sur un critère cognitif bien que le déclin cognitif soit la principale cause de désinsertion sociale des patients. Nous ne mentionnerons pas ici les aspects relatifs à l'utilisation de marqueurs ou biomarqueurs dits de substitution (« *surrogate biomarker* ») qui permettent de raccourcir la durée d'une étude en substituant le critère de jugement principal mesurable à long terme par un critère qui lui est corrélé et mesurable à court terme.

Actuellement, le plan d'expérience le plus utilisé dans la maladie de Huntington est le plan parallèle (Annexe A.2, Table 9). Dans le cas de petits effectifs, d'autres plans expérimentaux peuvent être préférés [41].

Parmi ces plans, nous pouvons citer le « *cross-over* » (Figure 5.B) où chaque patient reçoit les deux traitements et où la randomisation définit l'ordre d'administration des traitements. De même, nous pouvons citer le plan « *N-of-1* » (Figure 5.C) où un seul patient reçoit plusieurs traitements [42, 43, 44, 21]. Ce plan expérimental est utilisé lorsque l'on cherche à déterminer le meilleur traitement au niveau individuel. Il peut devenir très intéressant lorsque qu'il n'existe que peu de cas connus dans le monde. Pour ces deux plans expérimentaux, une période de latence (« *wash-out* ») entre les traitements est nécessaire afin que les effets du second traitement ne soient pas influencés par les effets du premier et que l'état du patient soit le même à chaque administration d'un traitement. Ces plans expérimentaux ne peuvent pas convenir dans les études de biothérapies dans la maladie

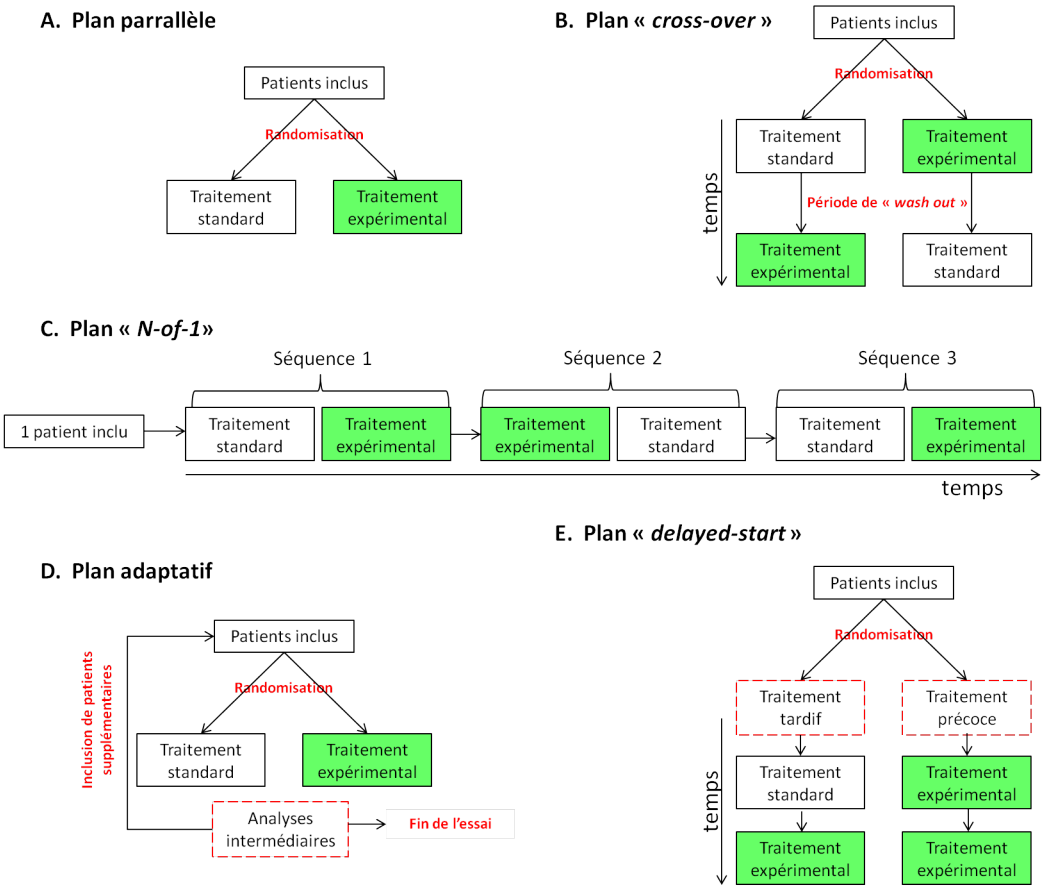


FIGURE 5 – Exemple de plans expérimentaux

de Huntington, car il s'agit principalement de traitements irréversibles, et même dans le cas où ils seraient réversibles, l'efficacité du traitement s'évaluant sur une période longue, le déclin naturel lié à la maladie ne garantit pas la comparabilité de l'état du patient à l'initiation de chaque traitement.

Les plans expérimentaux adaptatifs (Figure 5.D) constituent une autre approche en utilisant les données accumulées au cours de l'essai dans le but d'apporter des modifications à l'essai en cours. Ainsi le nombre de patients à inclure dans chaque bras de traitement peut varier au cours de l'essai. Cela peut induire des groupes déséquilibrés et donc réduire la puissance de l'étude, mais le nombre final de patients à inclure est souvent bien inférieur à celui nécessaire dans les essais en parallèle classique en particulier dans le groupe recevant le traitement le moins efficace [45, 46, 47]. Cette stratégie requiert des contraintes statistiques strictes comme par exemple le contrôle du risque α lors des analyses intermédiaires. De plus, il nécessite une réponse rapide de l'efficacité du traitement, ce qui est incompatible avec les greffes de cellules dans la maladie de Huntington.

Dans un plan d'expérience en « *delayed-start* » (Figure 5.E), l'essai clinique possède deux phases [48, 49, 50]. La première phase correspond à un plan parallèle classique où

les patients sont randomisés entre le nouveau traitement et le traitement de référence ou le placebo. Lors de la seconde phase, les deux groupes de patients reçoivent le nouveau traitement. Ce plan expérimental permet d'évaluer l'effet du traitement à la fois sur les symptômes mais aussi sur l'évolution de la maladie. Il permet de donner à tous les patients inclus dans l'essai le traitement à l'étude, la seconde phase se faisant forcément en levée de l'aveugle. Ce plan expérimental est celui qui a été utilisé dans MIG-HD et semble être le plus approprié actuellement pour un essai greffe dans la maladie de Huntington. L'ajout d'une phase de pré-randomisation dans MIG-HD a permis de doubler le nombre de patients dans une comparaison avant/après traitement.

Tous les plans expérimentaux cités ci-dessus (Figure 5) font l'hypothèse d'un effet du traitement identique chez tous les patients et ne tiennent pas compte d'une possible interaction entre l'efficacité du traitement et un facteur individuel. De nouveaux plans expérimentaux basés sur des principes de médecine stratifiée émergent. Ils supposent que l'on connaisse des marqueurs prédictifs de l'efficacité du traitement et testent les stratégies associées au fait de ne donner le traitement qu'aux patients possédant ce marqueur. La possible mise en place de ces plans stratifiés dans de futurs essais nécessite de les étudier et de les comparer dans le cas de petits effectifs. Nous avons donc identifié les principaux plans expérimentaux intégrant un marqueur prédictif. Outre le calcul du nombre de sujets nécessaires, ce travail a pour but l'identification des conséquences d'une mauvaise utilisation de ces plans expérimentaux et des paramètres diminuant la puissance en présence de petits effectifs.

En plus des facteurs prédictifs de l'efficacité du traitement, les facteurs pronostiques de la vitesse d'évolution de la maladie induisent une hétérogénéité supplémentaire. Cette hétérogénéité peut conduire à des différences de déclin naturel entre les groupes de patients d'un essai clinique, et cela malgré la randomisation, du fait des petits effectifs. Si les patients du groupe greffé ont une progression plus rapide de la maladie, cela peut conduire à une conclusion erronée de non-efficacité du traitement ; ou à une fausse efficacité du traitement dans le cas contraire [51]. La randomisation stratifiée sur des facteurs prédictifs du déclin permet de garantir un déclin naturel similaire entre les deux groupes de randomisation et ainsi mesurer l'effet réel du traitement [52]. Ainsi, nous avons évalué le déclin naturel des patients en fonction d'un polymorphisme génétique (le polymorphisme $Val^{158}Met$ sur le gène COMT). Ce travail a pour but de mettre en avant un nouveau marqueur pronostique du déclin chez les patients atteints de la maladie de Huntington.

Enfin, les critères cognitifs sont peu utilisés dans les essais cliniques de par la difficulté de mesure du déclin cognitif avec les échelles actuelles. L'une des hypothèses qui a été formulée est que l'effet « retest », c'est-à-dire l'amélioration des performances entre la première et la seconde passation du test empêche de mesurer le déclin cognitif [53]. Le

plan expérimental de MIG-HD propose d'utiliser une double visite à l'inclusion (avec un mois d'écart) pour homogénéiser les performances des patients en les familiarisant avec les tests (Figure 2). En effet, certains patients auront déjà passé ces tests plusieurs fois avant d'être inclus dans l'étude tandis que d'autres les découvriront pour la première fois lors de la visite d'inclusion. Grâce à ce plan particulier, nous pouvons d'une part évaluer l'effet retest sur un mois pour différentes échelles évaluant les troubles moteurs, fonctionnels, psychiatriques et cognitifs, et d'autre part évaluer l'impact de ce plan sur le mise en évidence d'un déclin cognitif chez les patients Huntington en un an. Ce travail a pour but d'identifier les conséquences de l'effet « retest » dans les essais cliniques et comment il peut être neutralisé de sorte à pouvoir utiliser les mesures cognitives dans les futurs essais de greffe pour la maladie de Huntington.

Description des chapitres

Notre travail a pour but d'une part de développer une méthode de clustering pour l'efficacité d'un traitement dans le cadre de données longitudinales et d'autres part de donner des axes d'amélioration pour les futurs essais de greffes dans la maladie de Huntington. Les analyses statistiques réalisées s'appuient sur les résultats obtenus dans l'essai MIG-HD ou dans la cohorte française des patients atteints de la maladie de Huntington (RHLF). Le manuscrit se divise en deux parties. La première concerne l'analyse de l'effet d'un traitement pour identifier des sous-groupes de patients répondeurs dans le cas de données longitudinales. Elle commence par un état de l'art de la modélisation des données longitudinales (Chapitre 1) et des méthodes de clustering (Chapitre 2). Puis nous y développons une méthode de clustering pour l'effet d'un traitement sur des données quantitatives continues longitudinales tenant compte de l'information pré-traitement (Chapitre 3).

La seconde partie concerne l'intégration marqueurs pronostique et/ou prédictif dans la conception des plans expérimentaux des essais cliniques. Elle commence par définir les marqueurs prédictifs et pronostiques (Chapitre 4). Nous démontrons que le polymorphisme $Val^{158}Met$ est un marqueur pronostique de la progression de la maladie de Huntington et montrons comment il peut être introduit dans les futurs essais cliniques (Chapitre 5). Puis nous comparons les plans expérimentaux intégrant un marqueur prédictif en terme de nombre de sujets nécessaires et d'impact de mauvaises connaissances des propriétés prédictives et pronostiques des marqueurs sur la puissance des études (Chapitre 6). Enfin, nous mettons en évidence l'effet « retest » dans les tests neuropsychologiques et montrons comment un plan expérimental adapté peut le neutraliser (Chapitre 7).

Première partie

Analyse de l'effet d'un traitement dans le cadre de données longitudinales et définition de sous-groupes répondeurs

Chapitre 1

Modélisation des données longitudinales (Etat de l'art)

Les données longitudinales sont des mesures d'une même variable, chez les mêmes patients, au cours du temps. Dans le cas des maladies neurodégénératives, il s'agit principalement de critères de jugement quantitatifs (scores), obtenus à l'aide de tests construits pour refléter l'évolution de la maladie (par exemple le score moteur de l'UHDRS dans la maladie de Huntington [54], le *Mini Mental Status Examination* (MMSE) dans la maladie d'Alzheimer [55], ...). Les effets plancher et plafond associés à ces tests ne permettent pas de considérer les scores comme des données linéaires et continues lorsque la période d'observation est longue (par exemple l'observation de l'évolution de la maladie dès l'apparition des premiers symptômes jusqu'au décès du patient). Cependant, dans le cas d'une durée d'observation courte, comme dans le cas d'un essai clinique, l'hypothèse d'une évolution linéaire des scores est admise. Cette hypothèse permet d'interpréter plus facilement les résultats et de réduire le nombre de paramètres à estimer, ce qui est important dans les études de petits effectifs.

La corrélation entre les mesures d'un même patient et la présence de données manquantes sont deux caractéristiques des données longitudinales. Le modèle à effets mixtes permet de considérer séparément les termes issus de la variabilité inter-patients de ceux issus de la variabilité intra-patient. L'algorithme d'espérance-maximisation (*Expectation-Maximization*, EM) permet d'estimer le modèle en présence de données incomplètes et par extension lorsque les patients ne sont pas suivis au même moment [56]. Nous posons ici les bases d'un modèle linéaire à effets mixtes et montrons comment un modèle mixte linéaire par morceaux avec deux pentes permet d'estimer l'effet d'un traitement sur l'évolution d'une maladie.

1.1 Le modèle linéaire à effets mixtes

1.1.1 Notations et modélisations

Soit le patient $i \in \{1,...,N\}$ et n_i le nombre d'observations pour le patient i. Soit y_{ij} le score du patient i mesuré au temps t_{ij}, où $j \in \{1,...,n_i\}$. La trajectoire du patient i est $Y_i = (y_{i1}...y_{in_i})$, le vecteur composé de toutes ses observations. Soit X_i et Z_i les matrices des covariables pour le patient i, de dimension $(n_i \times p)$ et $(n_i \times q)$, associées respectivement aux effets de population (p effets fixes) et aux effets individuels (q effets aléatoires). De manière générale, le modèle linéaire mixte s'écrit [57] :

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \tag{1.1}$$

où $\beta \in \mathbb{R}^p$ est appelé vecteur des effets fixes et $b_i \in \mathbb{R}^q$ vecteur des effets aléatoires spécifique au patient i. Les hypothèses de ce modèle sont $b_i \sim \mathcal{N}(0,D)$, $\varepsilon_{ij} \sim \mathcal{N}(0,\Sigma_i)$ ($\mathcal{N}$ représente la distribution de la loi normale) et $b_1,...,b_N,\varepsilon_1,...,\varepsilon_N$ indépendants. Conditionnellement à b_i, le vecteur Y_i suit une loi normale d'espérance $X_i\beta + Z_i b_i$ et de matrice de variance-covariance Σ_i.

Dans le cas où l'on s'intéresse uniquement à l'effet du temps, le modèle peut s'écrire :

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t_{ij} + \varepsilon_{ij} \tag{1.2}$$

où β_0 représente le score moyen au début de l'étude ($t = 0$, baseline), β_1 la pente d'évolution moyenne, b_{0i} l'effet individuel par rapport au score à baseline et b_{1i} l'effet individuel par rapport à la pente d'évolution moyenne. Les variances des effets individuels permettent de tenir compte de l'hétérogénéité des patients en terme d'évolution mais aussi en terme de performances à l'inclusion. Enfin le terme ε_{ij} permet de tenir compte de la variabilité intra-patient.

1.1.2 Estimation des paramètres du modèle marginal

Le modèle marginal s'écrit :

$$Y_i \sim \mathcal{N}(X_i\beta,\, Z_i D Z_i^T + \Sigma_i) \tag{1.3}$$

Vraisemblance du modèle

Notons R la matrice de variance-covariance du modèle marginal, matrice diagonale par blocs des R_1,...,R_N. Alors $R = ZDZ^T + \Sigma$ où $Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix}$ et Σ est la matrice diagonale par blocs des Σ_1,...,Σ_N. Lorsque les paramètres de variance, notés θ sont inconnus, la matrice R dépend de θ. Nous la notons $R(\theta)$. La vraisemblance du modèle s'écrit :

$$L(\beta,\theta;Y) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{n_i}{2}}|R_i(\theta)|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(Y_i - X_i\beta)^T R_i(\theta)^{-1}(Y_i - X_i\beta)\right\} \tag{1.4}$$

et la log-vraisemblance s'écrit :

$$l(\beta,\theta;Y) = -\frac{n}{2}log(2\pi) - \frac{1}{2}log|R(\theta)| - \frac{1}{2}(Y - X\beta)^T R(\theta)^{-1}(Y - X\beta) \tag{1.5}$$

avec $|R(\theta)|$ le déterminant de la matrice $R(\theta)$ et $n = \sum_{i=1}^{N} n_i$. Soit $\hat{\beta}$ l'estimateur de β par maximum de vraisemblance, alors $\hat{\beta}$ vérifie l'équation :

$$\frac{\partial l(\beta,\theta;Y)}{\partial \beta} = 0 \tag{1.6}$$

ce qui est équivalent à :

$$\hat{\beta} = (X^T R(\theta)^{-1} X)^{-1} X^T R(\theta)^{-1} Y \tag{1.7}$$

avec $\hat{\beta} = \hat{\beta}(\theta)$. Estimer β nécessite de connaître $R(\theta)$ et donc d'estimer les paramètres de variance θ. Deux méthodes permettent d'estimer θ : la méthode du maximum de vraisemblance (*Maximum Likelihood*, ML) et la méthode du maximum de vraisemblance restreinte (*REstricted Maximum Likelihood*, REML).

Estimation de θ par la méthode ML

La méthode ML consiste à estimer θ par le paramètre maximisant la vraisemblance (1.4) c'est-à-dire vérifiant $\frac{\partial l(\beta,\theta;Y)}{\partial \theta} = 0$ où β est remplacé par (1.7). Cette méthode pose un biais (sous-estimation) dans l'estimation des variances des effets aléatoires. En effet elle ne tient pas compte des degrés de liberté perdus dans l'estimation de β. Ce problème est corrigé dans la méthode REML.

Estimation de θ par la méthode REML

Soit r le rang de la matrice X. La méthode REML estime θ non plus en maximisant la vraisemblance des données Y (1.4) mais la vraisemblance restreinte qui correspond à la

vraisemblance des données définies par A^TY où A est une matrice de dimension $(n \times n-r)$ telle que $A^TX = 0$. Si $Y \sim \mathcal{N}(X\beta,R)$, alors $A^TY \sim \mathcal{N}(0,A^TRA)$. Ainsi la vraisemblance restreinte ne dépend plus de β.

Les méthodes ML et REML aboutissent à des équations non linéaires en θ pour lesquelles il n'existe pas de solution simple. Des algorithmes itératifs, tel que l'algorithme EM sont alors utilisés pour résoudre ces équations et estimer θ.

1.1.3 Estimation des effets aléatoires

En ce qui concerne les effets aléatoires, leur variance est souvent le paramètre d'intérêt. Mais il est possible de prédire les valeurs des effets aléatoires de chaque niveau. Les sources d'information permettant de prédire l'effet aléatoire associé au patient i sont les données observées pour le patient i ainsi que les paramètres de la distribution dont sont issus les effets aléatoires à savoir $\mathcal{N}(0,D)$. Les effets aléatoires sont prédits par :

$$\hat{b} = DZ^T(ZDZ^T + \Sigma)^{-1}(Y - X\beta) \tag{1.8}$$

1.2 Le modèle à effets mixtes linéaire par morceaux pour modéliser l'effet d'un traitement sur l'évolution de la maladie

1.2.1 Notations et modélisations

L'effet d'un traitement sur l'évolution d'un score continu peut se modéliser par un changement de pente. Le modèle linéaire par morceaux estime une première pente correspondant à la pente d'évolution pré-traitement et une seconde pente correspondant à la pente d'évolution post-traitement. Nous simplifions le modèle polynomial par morceaux proposé par Madsen et al. [58] pour obtenir le modèle linéaire par morceaux ci-dessous :

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t_{ij} + (\beta_2 + b_{2i}) \times (t_{ij} - \tau) \times \mathbb{1}(t_{ij} \geq \tau) + \varepsilon_{ij} \tag{1.9}$$

où y_{ij} correspond à la $j$ème observation du patient i au temps t_{ij} et τ représente le temps de début du traitement, commun à tous les patients. $\beta = (\beta_0,\beta_1,\beta_2)$ est le vecteur des effets fixes et $b_i = (b_{0i},b_{1i},b_{2i})$ est le vecteur des effets aléatoires spécifiques au patient i tel que $b_i \sim \mathcal{N}(0,D)$. Enfin, $\varepsilon_{ij} \sim \mathcal{N}(0,\Sigma_i)$ correspond à la variabilité des mesures. Les

interprétations des paramètres β_0, β_1, b_0 et b_1 restent les mêmes que pour l'équation (1.2). Les paramètres β_2 et b_2 sont associés au différentiel de pentes pré- et post-traitement avec β_2 le différentiel moyen et b_{2i} l'effet individuel du patient i. La variance associée à b_2 représente l'hétérogénéité de l'effet du traitement autour de l'effet moyen.

1.2.2 Application à MIG-HD

Dans l'essai clinique MIG-HD, les patients sont tous suivis avant et après traitement mais ne sont pas tous greffés au même moment. Les patients du groupe « greffe précoce » sont greffés après 13 mois de suivi tandis que les patients du groupe « greffe tardive » sont greffés après 33 mois de suivi. Nous proposons d'aligner les données des deux groupes sur la date de la première greffe qui deviendra le temps $t = 0$ (baseline) comme le montre la figure 6.

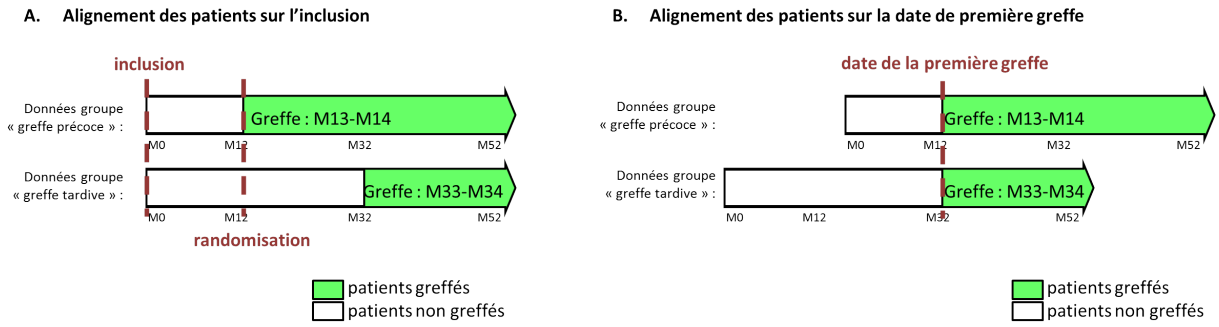


FIGURE 6 – Réalignement des données des groupes « greffe précoce » et « greffe tardive »

A. Alignement sur la date d'inclusion ; B. Alignement sur la date de la première greffe. En alignant les données sur la date de la première greffe, les données pré-greffe seront majoritairement du groupe « greffe tardive » tandis que les données post-greffe seront majoritairement du groupe « greffe précoce ».

En adaptant le modèle (1.9) à ces données, nous modifions l'interprétation des paramètres. Le tableau 1 résume l'interprétation des paramètres du modèle (1.9) dans le cas standard développé par Madsen et al. et dans le cas de son adaptation aux données de MIG-HD.

De façon plus générale, ce modèle peut être utilisé sur des données issues du bras de randomisation « traitement » d'un essai clinique longitudinal avec une période d'observation pré-traitement ou des deux bras de randomisation dans le cas d'un essai clinique « *delayed-start* ». Il peut aussi s'appuyer sur des données observationnelles de suivi de cohorte.

Lors de l'analyse de l'essai MIG-HD, nous avons appliqué ce modèle pour évaluer l'effet du traitement sur la pente d'évolution du score moteur de l'UHDRS. Ce modèle nous a

TABLE 1 – Interprétation des paramètres du modèle (1.9)

	Modèle présenté par Madsen et al.	Modèle adapté à MIG-HD
Interprétations modifiées :		
t	temps depuis l'inclusion	temps depuis la première greffe (les temps pré-greffe sont négatifs)
τ	délai entre l'inclusion et le traitement	délai entre l'initiation du traitement et l'effet du traitement
β_0	score moyen à l'inclusion	score moyen à l'initiation du traitement
Interprétations non modifiées :		
β_1	pente moyenne pré-traitement	pente moyenne pré-traitement
β_2	différentiel de pente moyen	différentiel de pente moyen

permis de tenir compte à la fois du plan expérimental en « *delayed-start* » et des temps de mesures décalés et/ou rajoutés suite à la difficulté de programmer les greffes. Lors de cette analyse, nous avons fait l'hypothèse d'un effet immédiat de la greffe, soit $\tau = 0$. Nous n'avons pas mis en évidence de différence de pente pré- et post-traitement dans cet essai.

Chapitre 2

Clustering des données quantitatives (Etat de l'art)

Le clustering est le terme générique désignant les méthodes d'apprentissage non supervisée permettant de construire des sous-groupes de données homogènes. Ces méthodes, où les sous-groupes ne sont pas pré-définis, se différencient des méthodes d'apprentissage supervisées qui visent à prédire des règles de classification à partir d'exemples de sous-groupes déjà définis. En général, les méthodes classiques de clustering s'appliquent sur des données transversales. Certaines études longitudinales de clustering s'intéressent aussi à des données transversales en résumant l'information par exemple par un coefficient de pente ou une durée de survie (méthode en deux étapes). D'autres, à l'inverse, vont utiliser les données longitudinales dans le clustering en y intégrant les mesures répétées. Parce qu'aucune de ces méthodes de clustering pour données longitudinales n'est satisfaisante dans notre cas, où l'on souhaite trouver des patients répondeurs à un traitement, nous avons proposé une nouvelle méthode. Notre méthode appartient à la catégorie des méthodes en deux étapes et sera décrite au chapitre 3. Parce que nous utiliserons des méthodes pour données quantitatives transversales et que nous comparerons nos résultats avec les méthodes pour données quantitatives longitudinales, nous faisons ici une revue de ces méthodes. Les algorithmes que nous décrivons appartiennent aux grands groupes de méthodes représentés sur la figure 7.

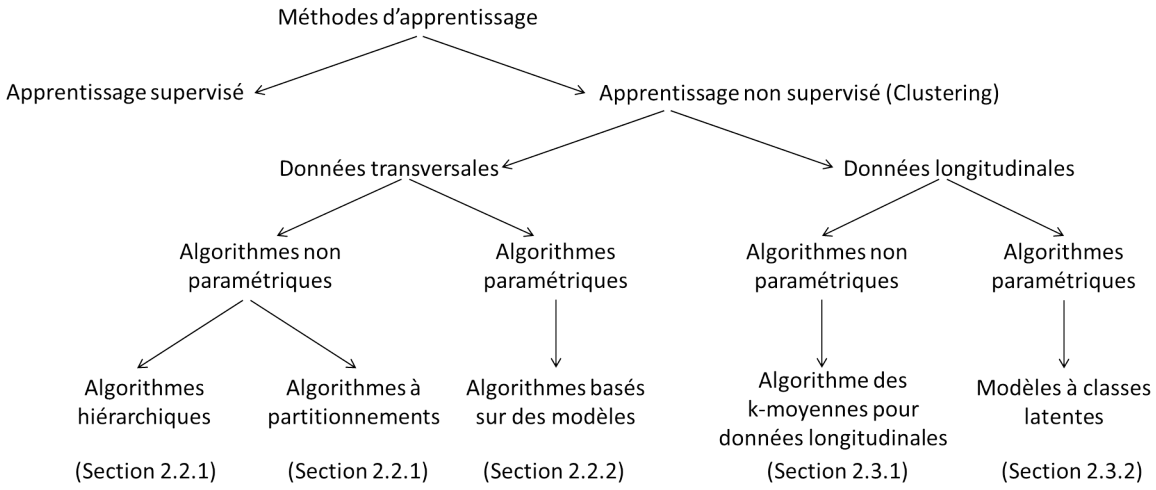


FIGURE 7 – Schématisation de la place des algorithmes que nous décrivons parmi les techniques d'apprentissage

2.1 Mesures de dissimilarité entre deux observations

Les algorithmes de clustering se basent sur des mesures de similarité (s) ou de dissimilarité (d) entre deux observations. Pour les algorithmes non paramétriques, d est une application, appelée distance, à valeurs dans l'ensemble des réels positifs ($d : E \times E \longrightarrow \mathbb{R}^+$) où $E \subset \mathbb{R}^p$ représente l'ensemble de nos observations. L'application d vérifie les propriétés :

- de symétrie : $\forall(i,j) \in E^2, d(i,j) = d(j,i)$;
- de séparation : $\forall(i,j) \in E^2, d(i,j) = 0 \Leftrightarrow i = j$;
- d'inégalité triangulaire : $\forall(i,j,m) \in E^3, d(i,j) \leq d(i,m) + d(m,j)$.

Soit x_i un vecteur de dimension p des caractéristiques de l'observation i où $x_{i1},...,x_{ip}$ sont des valeurs quantitatives continues (scores). La distance entre deux observations i et j se calcule à partir de ces scores. La distance est d'autant plus faible que les scores sont proches. La table 2 présente les principales distances pouvant être utilisées au sein des algorithmes non paramétriques [59, 60].

La distance euclidienne, aussi appelée norme L_2, est la distance la plus connue. Elle peut être assimilée à la distance parcourue à vol d'oiseau entre deux points A et B, tandis que la distance de Manhattan [61], ou norme L_1, serait assimilée à la distance parcourue en marchant dans des rues suivant un quadrillage. La distance de Chebyshev, ou norme $L_{+\infty}$, correspond à la plus grande projection de B sur les axes de l'espace dont le centre serait défini par A (voir Figure 8 pour un exemple en dimension 2). Les distances euclidienne, de Manhattan et de Chebyshev sont des cas particuliers de la distance de Minkowski avec respectivement le paramètre de Minkowski r égal à 1, 2 et $+\infty$. Lorsque $p = 1$ la distance de Minkowski est la valeur absolue de la différence de scores entre les deux observations

TABLE 2 – Exemples de distances pouvant être utilisées au sein des algorithmes non paramétriques pour les données quantitatives

Distance	$d(i,j)$	**Distance**	$d(i,j)$
Minkowski	$\left(\sum_{\ell=1}^{p} \mid x_{\ell,i} - x_{\ell,j} \mid^r\right)^{1/r}$ $r \geq 1$	Euclidienne	$\sqrt{\sum_{\ell=1}^{p} \mid x_{\ell,i} - x_{\ell,j} \mid^2}$
Canberra	$\sum_{\ell=1}^{p} w_\ell(i,j) \mid x_{\ell,i} - x_{\ell,j} \mid$	Manhattan	$\sum_{\ell=1}^{p} \mid x_{\ell,i} - x_{\ell,j} \mid$
	$w_\ell(i,j) = \begin{cases} 0 \quad \text{si} \quad x_{\ell,i} = x_{\ell,j} = 0 \\ \frac{1}{\mid x_{\ell,i} \mid + \mid x_{\ell,j} \mid} \quad \text{sinon} \end{cases}$	Chebyshev	$\max_{\ell \in 1..p} \mid x_{\ell,i} - x_{\ell,j} \mid$
Pearson ($p > 1$)	$1 - \frac{\sum_{\ell=1}^{p} x_{\ell,i} x_{\ell,j}}{\sqrt{\sum_{\ell=1}^{p} x_{\ell,i}^2 \sum_{\ell=1}^{p} x_{\ell,j}^2}}$	Corrélation ($p > 1$)	$1 - \frac{cov(x_i,x_j)}{\sqrt{var(x_i) var(x_j)}}$

$d(i,j)$ est la distance entre deux observations i et j lorsque $x_{\ell,i}$ et $x_{\ell,j}$ représentent les scores à la caractéristique ℓ avec $\ell \in \{1,...,p\}$. Les distances de Pearson et de Corrélation ne sont pas définies pour $p = 1$. Les formules présentées dans ce tableau n'utilisent pas de terme de pondération, supposant que le même poids est donné à toutes les caractéristiques $\ell \in \{1,...,p\}$.

quel que soit r. Les distances euclidienne, de Manhattan et de Chebyshev sont donc égales dans le cas $p = 1$.

La distance de Canberra [62] peut être vue comme une version pondérée de la distance de Manhattan. Le dénominateur assure une forte sensibilité aux faibles variations lorsque (X_i,X_j) est proche de (0,0). Pour $p = 1$, cette distance est à valeur dans $[0;1]$ où 1 est atteint dès que X_i et X_j sont de signes opposés.
La distance de Pearson, aussi appelée distance de Pearson non centrée, ou distance angulaire est égale à $1 - cos(\theta)$ où θ est l'angle entre les deux vecteurs X_i et X_j. Cette distance est à valeur dans [0,2]. La distance de corrélation, aussi appelée distance de Pearson centrée est à valeur dans [0,2].

Différentes distances appliquées sur les mêmes données dans le but de construire des sous-groupes homogènes peuvent conduire à des résultats différentes. Il n'y a pas une mesure qui soit optimale par rapport aux autres, sauf peut-être d'utiliser celle qui semble donner la meilleure interprétation [63].

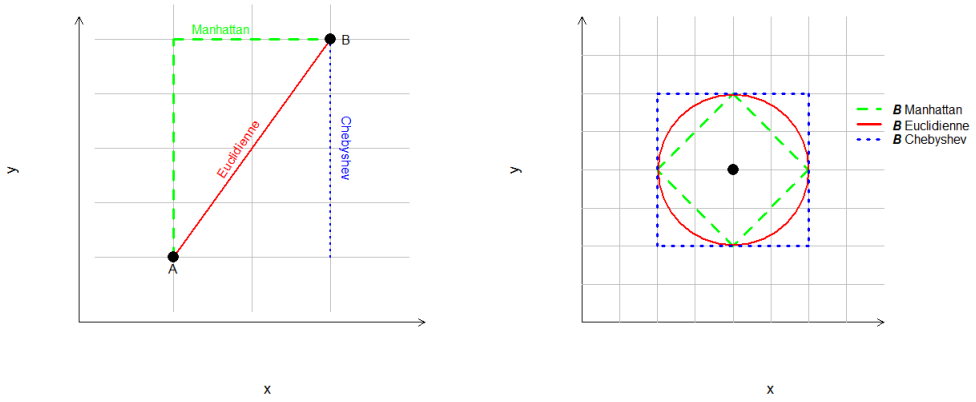


FIGURE 8 – Schématisation des distances euclidienne, de Manhattan et de Chebyshev dans un espace de dimension 2

Soit z une distance fixée, A un point et $\mathcal{B}$ l'ensemble des points B tel que $d(A,B) = z$. Alors, en dimension 2, $\mathcal{B}$ est défini par un cercle pour la distance euclidienne, un losange dont les sommets sont sur les axes qui définissent l'espace centré en A pour la distance de Manhattan ou un carré dont les centres des côtés sont sur les axes qui définissent l'espace centré en A pour la distance de Chebyshev. A noter qu'en dimension 1, $\mathcal{B}$ est défini par les bornes d'un segment pour les trois distances et en dimension 3, $\mathcal{B}$ est défini par une sphère (distance euclidienne), un octoèdre (distance de Manhattan) ou un cube (distance de Chebyshev).

2.2 Cas des données transversales

2.2.1 Algorithmes non paramétriques

Les principaux algorithmes non paramétriques se divisent en deux familles : les algorithmes à partitionnement et les algorithmes hiérarchiques.

• Les algorithmes à partitionnement centroïdes

Soit $\mathcal{O}$ un ensemble de n observations chacune déterminée par un vecteur $x_i = (x_{i1}...x_{ip}) \in \mathbb{R}^p$. Les algorithmes à partitionnement rigide divisent $\mathcal{O}$ en un nombre prédéfini K de clusters $\mathcal{C} = (\mathcal{C}_1,...,\mathcal{C}_K)$, aussi appelés partitions. Chaque observation de $\mathcal{O}$ appartient à un et un seul cluster (c'est-à-dire $\forall(k,k') \in \{1..K\} \setminus k \neq k', \mathcal{C}_k \bigcap \mathcal{C}_{k'} = \emptyset$ et $\bigcup_{k=1}^{K} \mathcal{C}_k = \mathcal{O}$). Chaque cluster $\mathcal{C}_k$ possède un centre appelé noyau et noté z_k. Les clusters sont d'abord déterminés aléatoirement puis redéfinis en attribuant chaque observation au cluster le plus proche (c'est-à-dire le cluster pour lequel la distance entre lui-même et le noyau est minimale) itérativement, jusqu'à la stabilité des clusters.

L'algorithme à partitionnement le plus connu est celui des K-moyennes [64, 65]. Dans cet algorithme le noyau est défini par la moyenne de tous les points appartenant au cluster ($\forall k \in \{1..K\}, z_k = \frac{1}{card(\mathcal{C}_k)} \sum_{i \in \mathcal{C}_k} x_i$). Les étapes de cet algorithme sont détaillées ci-dessous.

1. Initialisation : $\ell = 0$
   Tirer aléatoirement K observations représentant les noyaux initiaux $z_1^{(\ell)},...,z_K^{(\ell)}$
2. Assigner chaque observation i au cluster le plus proche :
   $\forall k, \mathcal{C}_k^{(\ell)} = \left\{ x_i \in \mathcal{O} \setminus k = \underset{k' \in 1..K}{\arg\min}(d(x_i, z_{k'}^{(\ell)})) \right\}$
3. Définir les nouveaux noyaux $z_k^{(\ell+1)}$ qui sont les moyennes des cluster $\mathcal{C}_k^{(\ell)}$:
   $\forall k, z_k^{(\ell+1)} = \frac{1}{card(\mathcal{C}_k^{(\ell)})} \sum_{x_i \in \mathcal{C}_k^{(\ell)}} x_i$
   Tant que $\exists\, k \in \{1..K\} \setminus z_k^{(\ell+1)} \neq z_k^{(\ell)}$, répéter les étapes 2 et 3.

Les étapes 2 et 3 sont répétées jusqu'à convergence de l'algorithme, c'est-à-dire jusqu'à ce que les clusters restent inchangés. La figure 9 illustre cet algorithme.

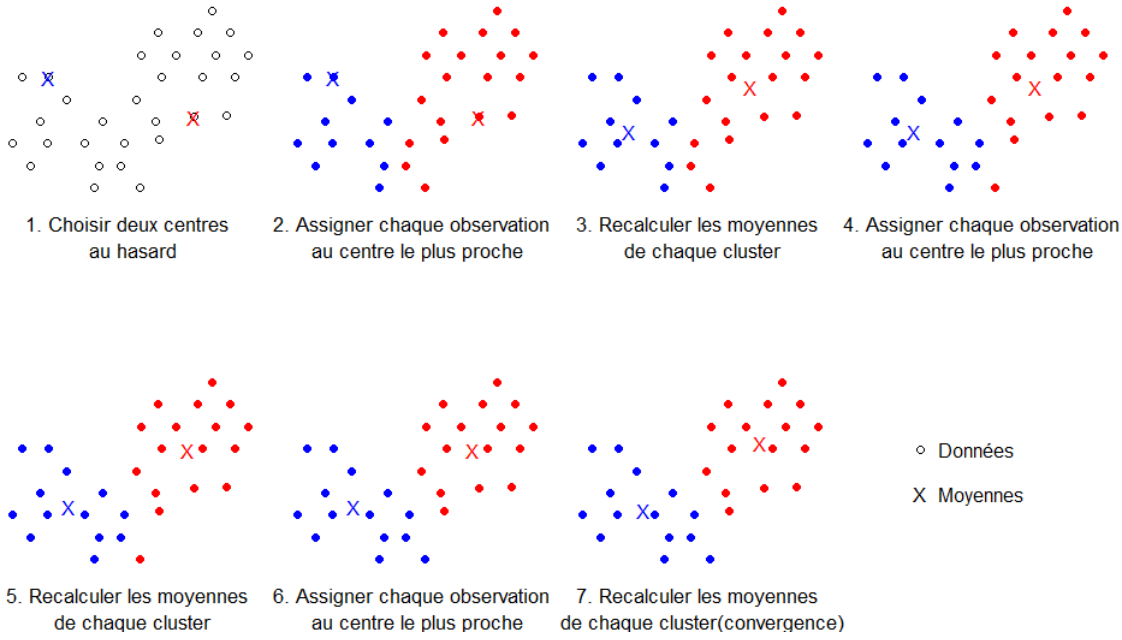


FIGURE 9 – L'algorithme des K-moyennes

Dans cet exemple chaque point représente une observation dans un espace de dimension 2. La distance utilisée pour la réalisation de cette figure est la distance euclidienne.

L'avantage de cette méthode réside dans sa simplicité et sa capacité à toujours converger [66]. Cependant, l'algorithme peut converger vers des *optima* locaux différents dès lors que les paramètres de l'initialisation changent et donc ne pas converger vers la meilleure classification. Tester plusieurs paramètres d'initialisation ou utiliser un algorithme permettant de définir les meilleurs paramètres d'initialisation peut pallier ce problème [67, 68]. De plus, l'algorithme des K-moyennes est très sensible aux valeurs aberrantes (« *out-*

liers »). L'impacte des valeurs extrêmes ou aberrantes peut être minimisé en remplaçant la moyenne par la médiane. L'algorithme à partitionnement K-médoïdes fonctionne sur le même principe que les K-moyennes à la différence que le noyau de chaque cluster n'est plus représenté par la moyenne du cluster mais par la donnée la plus centrale (médoïde) du cluster [69]. La médoïde z_k du $k$ème cluster est définie par $z_k = \underset{x_j \in \mathcal{C}_k}{\arg\min} \sum_{x_i \in \mathcal{C}_k} d(x_j,x_i)$. La médoïde peut être assimilée à une médiane. Ainsi, cet algorithme est considéré comme plus robuste que l'algorithme des K-moyennes [70]. De plus, cet algorithme reste efficace même dans le cas de petits effectifs.

Ces algorithmes peuvent s'étendre à des algorithmes à frontières floues autorisant une observation à appartenir à plusieurs clusters avec un certain degré d'appartenance [71]. Si on note $c_{i,k}$ le degré d'appartenance de x_i au cluster $\mathcal{C}_k$, alors $0 \leq c_{i,k} \leq 1$ et $\sum_{k=1}^{K} c_{i,k} = 1$. Les algorithmes à partitionnement rigide sont un cas particulier où $c_{i,k}$ ne peut prendre que les valeurs 0 ou 1. L'avantage des algorithmes à frontières floues réside dans le fait qu'on peut combiner l'information donnée par $c_{i,k}$ avec les *a priori* sur les observations pour déterminer le meilleur cluster auquel assigner chaque observation.

Toutes ces méthodes nécessitent de connaître le nombre de clusters et ne peuvent pas identifier des groupes non convexes.

• Les algorithmes à partitionnement par densité

Les algorithmes à densité ont émergé à la fin des années 90 et reposent sur le principe qu'un cluster est formé de nombreuses observations très proches et que les observations isolées sont des données aberrantes [72]. Ainsi, toutes les observations n'appartiennent pas forcément à un cluster $\left(\bigcup_{k=1}^{K} \mathcal{C}_k \subseteq \mathcal{O}\right)$. L'algorithme visite les données une par une en leur appliquant les règles suivantes, où q et ε sont des paramètres prédéfinis par l'utilisateur :

1. une observation $x \in \mathcal{O}$ peut définir un cluster si au moins q observations sont à une distance inférieure à ε de x.
2. une observation $y_p \in \mathcal{O}$ appartient à un cluster si elle peut le définir ou si le cluster lui est accessible, c'est-à-dire s'il existe un chemin $y_1,...,y_p$ tel que y_1 peut définir un cluster et que $\forall i = 1..p-1$, la distance entre y_i et y_{i+1} est inférieure à ε.
3. une observation z est une donnée aberrante si elle ne peut ni définir un cluster ni avoir un cluster accessible.

La figure 10 schématise ces définitions.

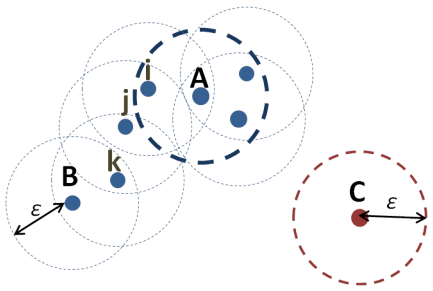


FIGURE 10 – Schématisation de l'algorithme à partitionnement par densité

Si $q = 3$, les observations bleues forment un cluster dont la seule observation qui le définit est l'observation A (seule observation avec au moins 3 observations à une distance inférieure à ε). L'observation B appartient à ce cluster car le cluster lui est accessible via la suite d'observations A-i-j-k-B. Enfin, l'observation C est considérée comme une donnée aberrante.

L'algorithme mis en place est le suivant :

Soit q le nombre minimal d'observations de $\mathcal{O}$ d'une distance inférieure à ε de $x \in \mathcal{O}$ nécessaires à la formation d'un cluster.

$\forall x \in \mathcal{O}$ n'ayant pas été visitée :

- Considérer x comme ayant été visitée.
- Soit $\mathcal{V}_x = \{y \in \mathcal{O} \setminus d(x,y) \leq \varepsilon\}$ le voisinage de x
  - Si $Card(\mathcal{V}_x) < q$, alors x est considérée comme une donnée aberrante.
  - Si $Card(\mathcal{V}_x) \geq q$, alors x appartient au cluster $\mathcal{C}_x$

    $\forall y \in \mathcal{V}_x$:
    - Si y n'a pas été visitée :
      - Soit $\mathcal{V}_y = \{z \in \mathcal{O} \setminus d(y,z) \leq \varepsilon\}$ le voisinage de y
      - Si $Card(\mathcal{V}_y) \geq q$, alors $\mathcal{V}_x \leftarrow \mathcal{V}_x \cup \mathcal{V}_y$:
    - Si y n'appartient à aucun cluster, alors $y \in \mathcal{C}_x$

L'un des avantages de cet algorithme est qu'il est capable d'identifier des structures de clusters non convexes. De plus, il définit lui-même le nombre de clusters. Cependant le nombre de clusters dépendra des paramètres choisis pour q et ε. Lorsque les clusters ont des densités différentes, cet algorithme donne de mauvais résultats.

• Les algorithmes hiérarchiques

Le Clustering Ascendant Hiérarchique (CAH) est un algorithme déterministe qui part d'un état où il y a n clusters, chacun étant une observation de $\mathcal{O}$ pour arriver à un état où il n'y a qu'un seul cluster $\mathcal{O}$. L'algorithme regroupe au fur et à mesure les deux clusters les plus proches jusqu'à n'en former qu'un seul, comme le montre le dendrogramme de la figure 11. Il crée une décomposition hiérarchique des observations. Contrairement

aux K-moyennes et K-médoïdes, cet algorithme ne suppose pas de nombre de clusters *a priori* [73].

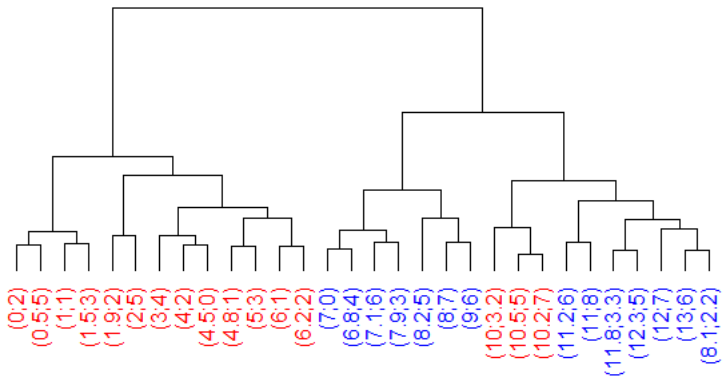


FIGURE 11 – Un dendrogramme, résultat de l'algorithme ascendant hiérarchique

Dans cet exemple chaque point représente une observation dans un espace de dimension 2 de coordonnées $(x;y)$. La distance utilisée pour cette figure est la distance euclidienne. Il s'agit des mêmes coordonnées utilisées dans la figure 9. Les couleurs rouge et bleue font référence aux clusters définis par l'algorithme des K-moyennes. Les nœuds du dendrogramme représentent les clusters tandis que la hauteur des branches représente la distance à laquelle les clusters sont fusionnés.

Les distances entre les clusters peuvent avoir plusieurs définitions. Dans la méthode dite « lien simple » (« *single linkage* »), ou technique du plus proche voisin, la distance entre deux clusters est la plus petite distance entre les observations de chaque cluster [74]. Mathématiquement, cela se traduit par $d(\mathcal{C},\mathcal{C}') = \min\{d(x_i,x_j), x_i \in \mathcal{C}, x_j \in \mathcal{C}'\}$. Dans la méthode dite « lien complet » (« *complete linkage* »), la distance entre deux clusters est la plus grande distance entre les observations de chaque cluster [75]. Mathématiquement, cela se traduit par $d(\mathcal{C},\mathcal{C}') = \max\{d(x_i,x_j), x_i \in \mathcal{C}, x_j \in \mathcal{C}'\}$. Ces deux méthodes ne tiennent pas compte de la structure des clusters. Dans la méthode dite centroïde, la distance entre deux clusters est la distance entre les moyennes de chaque cluster. Mathématiquement, cela se traduit par $d(\mathcal{C},\mathcal{C}') = d(\bar{x},\bar{x'})$ avec $\bar{x} = \frac{1}{card(\mathcal{C})} \sum_{x_i \in \mathcal{C}} x_i$ et $\bar{x'} = \frac{1}{card(\mathcal{C}')} \sum_{x_j \in \mathcal{C}'} x_j$. Dans cette méthode, la fusion de deux clusters est dominée par le cluster ayant le plus d'observations [76]. Enfin, la méthode dite « lien moyen » (« *average linkage* »), définit la distance entre deux clusters comme la moyenne des distances entre chaque observations des clusters. Mathématiquement, cela se traduit par $d(\mathcal{C},\mathcal{C}') = \frac{1}{card(\mathcal{D})} \sum_{d \in \mathcal{D}} d$ où $\mathcal{D} = \{d(x_i,x_j), x_i \in \mathcal{C}, x_j \in \mathcal{C}'\}$. Cette méthode semble être la plus robuste. Ces définitions sont représentées sur la figure 12.

A la suite de ces méthodes, plusieurs extensions ont émergé, comme par exemple l'utilisation de la valeur médiane ou d'un système de pondération [77, 78, 79]. Toutes les méthodes développées pour le CAH peuvent se transposer au clustering descendant hiérarchique. L'algorithme est similaire à l'exception qu'il part d'un état où il n'y a qu'un

seul cluster $\mathcal{O}$ qui regroupe toutes les données, pour arriver à un état où chaque observation de $\mathcal{O}$ représente un cluster. L'algorithme dissocie au fur et à mesure les deux clusters les plus éloignés.

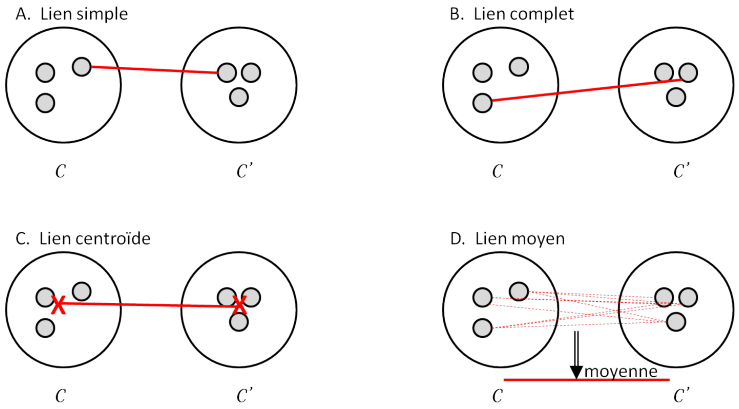


FIGURE 12 – Représentation schématique des définitions des distances ente deux clusters utilisées dans l'algorithme ascendant hiérarchique

2.2.2 Algorithmes paramétriques

Les algorithmes paramétriques se basent sur l'écriture de modèles dont peuvent être issues les observations de l'ensemble $\mathcal{O}$. Par rapport aux algorithmes non paramétriques, ceux-ci ont l'avantage de pouvoir réaliser des inférences et des tests statistiques sur $\mathcal{O}$. Certains algorithmes non paramétriques peuvent être vus comme des approximations d'algorithmes paramétriques.

• Les modèles de mélange

Les modèles de mélange fini considèrent que les observations $x_i \in \mathcal{O}$ forment des clusters chacun ayant une distribution de probabilité différente. Les distributions peuvent ne pas appartenir à la même famille, ou appartenir à la même famille mais différer dans les valeurs des paramètres. La densité (f) de la loi de probabilité dont sont issues les observations s'écrit :

$$f(x;\pi,\theta) = \sum_{k=1}^{K} \pi_k f_k(x,\theta_k) \tag{2.1}$$

où K est le nombre de clusters, f_k est la densité de la loi de probabilité pour les observations du cluster $\mathcal{C}_k$ et π_k est la probabilité pour qu'une observation appartienne au cluster $\mathcal{C}_k$ et vérifie $\sum_{k=1}^{K} \pi_k = 1$.

Nous développons ici le cas d'un mélange de lois gaussiennes. En dimension 1, $\theta_k = (\mu_k;\sigma_k)$ où μ_k est la moyenne du cluster $\mathcal{C}_k$ et σ_k son écart-type, tel que $\forall i,k \setminus x_i \in \mathcal{C}_k, \quad x_i \sim \mathcal{N}(\mu_k,\sigma_k^2)$. Les hypothèses considérées sont l'égalité ou l'inégalité des variances

entre chaque cluster. En dimension ≥ 2, $\theta_k = (\mu_k; \Sigma_k)$ où μ_k et Σ_k sont la moyenne et la matrice de variance-covariance associées aux observations du cluster $\mathcal{C}_k$. La densité f_k s'écrit alors :

$$f_k(x; \theta_k) = \frac{1}{\sqrt{det(2\pi\Sigma_k)}} exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right\} \tag{2.2}$$

La matrice de variance-covariance peut s'écrire sous la forme :

$$\Sigma_k = \lambda_k Q_k D_k Q_k^{-1} \tag{2.3}$$

où Q_k est la matrice inversible des vecteurs propres de Σ_k et $\lambda_k D_k$ est la matrice diagonale des valeurs propres de Σ_k tel que $det(D_k) = 1$. Alors $\lambda_k = \sqrt[n]{det(\Sigma_k)}$ où $det(\Sigma_k)$ est le produit des valeurs propres de Σ_k [80]. L'écriture sous la forme 2.3 permet de définir un volume (λ_k), une forme (D_k) et une orientation (Q_k). La figure 13 représente les données (x,y) dans un espace de dimension 2 pour différentes hypothèses sur λ_k et D_k où $k \in \{1,2\}$ [81].

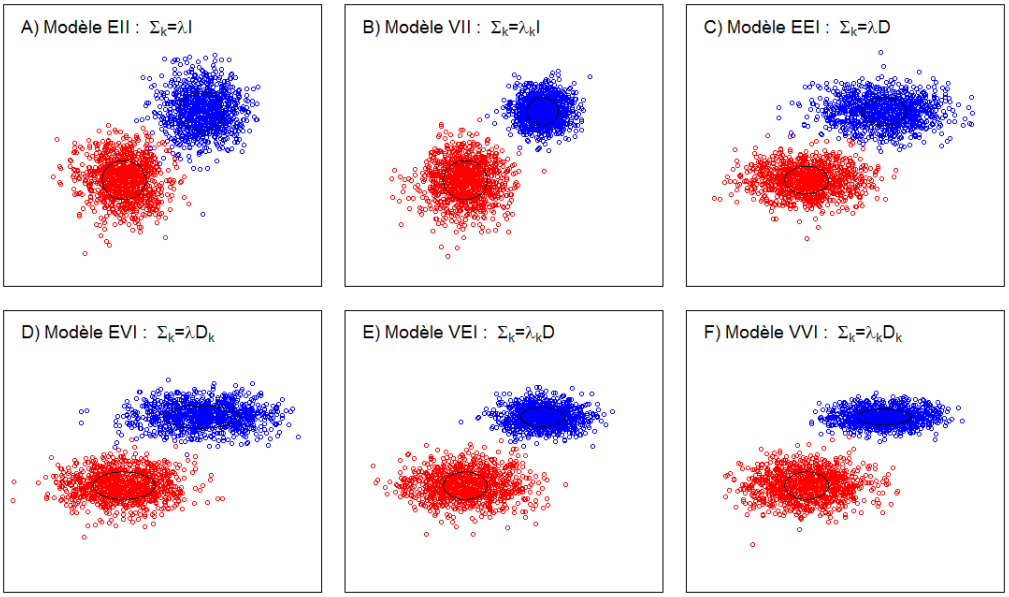


FIGURE 13 – Représentation schématique de données dans un espace de dimension 2 pour $k \in \{1,2\}$, $Q_k = I$ et différentes hypothèses sur λ_k et D_k

Les points rouges correspondent aux observations du cluster 1 et les points bleus aux observations du cluster 2. Les ellipses noires correspondent aux ellipses d'équidensités (de grand rayon $\sqrt{\lambda_k D_k[1,1]}$ et de petit rayon $\sqrt{\lambda_k D_k[2,2]}$) associées à la matrice de variance covariance de chaque cluster. Ces modèles s'écrivent sous la forme Volume (E=égal ; V=variable)/ Forme (E=égale ; V=variable ; I=sphérique)/ Direction (E=égale ; V=variable ; I=Parallèle à un axe). Tous ces modèles correspondent à des cas où les variables x et y sont non corrélés (matrice Σ_k diagonale) d'où Direction=I. Dans les cas A et B, les variances de x et y sont égales au sein de chaque cluster avec en plus égalité entre les clusters pour le cas A.

L'estimation des paramètres se fait via l'algorithme EM pour estimer le maximum de vraisemblance en présence de données incomplètes. Les observations x_i sont supposées incomplètes et les données complètes sont le couple (x_i,z_i) où z_i est un vecteur de longueur K tel que : $z_{ik} = 1$ si $x_i \in \mathcal{C}_k$ et $z_{ik} = 0$ sinon. Ainsi, la fonction de densité pour les données complètes est :

$$f(x;\pi,\theta) = \sum_{k=1}^{K} \pi_k \left(\frac{1}{\sqrt{det(2\pi\Sigma_k)}} exp\left\{ -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) \right\} \right)^{z_{ik}} \tag{2.4}$$

L'algorithme EM a la même forme quel que soit le modèle considéré. Seul l'estimation de Σ_k change [82]. Par exemple, pour le modèle VII ($\Sigma_k = \lambda_k I$), l'algorithme EM se résume comme ci-après (où les étapes M et E se répètent jusqu'à convergence), où n est le nombre d'observations dans un espace de dimension p.

- Initialisation : $\hat{z}_{ik}$
- **Etape M : maximisation de la vraisemblance, connaissant** $\hat{z}_{ik}$

  $\hat{n}_k \leftarrow \sum_{i=1}^{n} \hat{z}_{ik}$

  $\hat{\pi}_k \leftarrow \frac{\hat{n}_k}{n}$

  $\hat{\mu}_k \leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ik} x_i}{\hat{n}_k}$

  $\hat{W}_k \leftarrow \sum_{i=1}^{n} z_{ik}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$

  $\hat{\lambda}_k \leftarrow \frac{tr(W_k)}{p \times n_k}$

  $\hat{\Sigma}_k \leftarrow \hat{\lambda}_k I$

- **Etape E : estimation de** $\hat{z}_{ik}$

  $\hat{z}_{ik} < -\frac{\hat{\pi}_k f_k(x_i;\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{j=1}^{K} \hat{\pi}_j f_j(x_i;\hat{\mu}_j,\hat{\Sigma}_j)}$

2.3 Cas des donnés longitudinales

Lorsque les observations sont des évolutions d'un score dans le temps, nous cherchons à construire des sous-groupes de trajectoires. Plusieurs algorithmes de clustering ont été développés pour les données longitudinales. Nous présentons ici deux algorithmes, l'un

non paramétrique et l'autre paramétrique.

2.3.1 Un algorithme non paramétrique : clustering des données longitudinales par K-moyennes

L'idée de l'algorithme KML (*K-Means for Longitudinal data*) est d'appliquer l'algorithme des K-moyennes aux observations $y_i = (y_{i1},...,y_{iT})$ où y_{it} est le score obtenu par le patient i au temps t. La distance entre deux patients i et j est calculée par la distance euclidienne ou la distance de Manhattan appliquée aux observations y_i et y_j [34]. Pour pallier le problème des données manquantes, les distances utilisent l'ajustement de Gower [83] ci-dessous :

$$w_{ijt} = \begin{cases} 1 & \text{si } y_{it} \text{ et } y_{jt} \text{ ne sont pas manquntes} \\ 0 & \text{sinon} \end{cases} \tag{2.5}$$

Distance euclidienne avec ajustement de Gower :

$$d(i,j) = \sqrt{\frac{1}{\sum_{t=1}^{T} w_{ijt}} \sum_{t=1}^{T} w_{ijt}(y_{it} - y_{jt})^2} \tag{2.6}$$

Distance de Manhattan avec ajustement de Gower :

$$d(i,j) = \frac{1}{\sum_{t=1}^{T} w_{ijt}} \sum_{t=1}^{T} w_{ijt}|y_{it} - y_{jt}| \tag{2.7}$$

Cet algorithme montre de bonnes performances en étude de simulation quelle que soit l'allure des trajectoires. Cette méthode peut utiliser toute autre distance et peut être extrapolée au cas d'une évolution conjointe de deux variables continues [84].

2.3.2 Un algorithme paramétrique : clustering des données longitudinales par modèle mixte à classes latentes

Le modèle LCMM (*Latent Class Mixed Model*) modélise la trajectoire des patients par un modèle linéaire mixte, en supposant que la trajectoire diffère d'un cluster à l'autre [35]. Cette méthode estime conjointement deux modèles qui sont :

- la trajectoire conditionnellement au cluster
- la probabilité d'appartenir à chaque cluster

Les trajectoires spécifiques et les probabilités d'appartenance à un cluster peuvent être modélisées en fonction de covariables.

Soit y_{ij} la $j$ème observation du patient i au temps t_{ij}. Alors y_i correspond à la trajectoire du patient i. Si le patient i appartient au cluster $\mathcal{C}_k$, l'évolution de y_{ij} se modélise conditionnellement au cluster :

$$y_{ij}|\mathcal{C}_k = (\beta_0^{(k)} + b_{0i}^{(k)}) + (\beta_1^{(k)} + b_{1i}^{(k)}) \times t_{ij} + \varepsilon_{ij} \tag{2.8}$$

où $(\beta_0^{(k)},\beta_1^{(k)})$ est le vecteur des effets fixes et $(b_0^{(k)},b_1^{(k)})$ celui des effets aléatoires avec $(b_{0i}^{(k)},b_{1i}^{(k)}) \sim \mathcal{N}(\begin{pmatrix} 0 \\ 0 \end{pmatrix},\Sigma_k)$. Σ_k représente la variance-covariance spécifique au cluster k. Enfin le terme d'erreurs est défini par $\varepsilon_{ij} \sim \mathcal{N}(0,\sigma_\varepsilon^2)$.

La probabilité pour le patient i d'appartenir au cluster k (π_{ik}) s'explique grâce aux covariables X_i par un modèle de régression logistique multinomial :

$$\pi_{ik} = \mathbb{P}(i \in \mathcal{C}_k|X_i) = \frac{exp(\alpha_0^{(k)} + X_i^T\alpha_1^{(k)})}{\sum_{\ell=1}^{K} exp(\alpha_0^{(\ell)} + X_i^T\alpha_1^{(\ell)})} \tag{2.9}$$

où $\alpha_0^{(k)}$ et $\alpha_1^{(k)}$ sont les coefficients de régressions associés à la régression logistique multinomiale. Soit K la classe de référence tel que $\alpha_0^{(K)} = \alpha_1^{(K)} = 0$.

L'estimation par maximum de vraisemblance peut se faire via l'algorithme EM [85] ou par la maximisation directe de la vraisemblance observée [86].

2.4 Estimation du nombre de clusters

Toutes les méthodes que nous avons décrites ci-dessus, à l'exception du modèle à partitionnement par densité, nécessitent de connaître à l'avance le nombre de clusters. Cependant, nous ne pouvons pas toujours connaître ce nombre. Il convient alors de considérer plusieurs nombres de clusters et de choisir le meilleur sur la base d'un critère. Pour les méthodes non paramétriques le critère le plus souvent utilisé est le critère de Calinski et Harabasz [87]. Pour les méthodes non paramétriques, le critère le plus utilisé est le critère d'information bayésien (BIC, *Bayesian Information Criterion*) [88, 89].

Chapitre 3

Méthode de clustering pour l'effet d'un traitement prenant en compte l'information pré-traitement dans le cadre de données longitudinales

Les méthodes de clustering pour données longitudinales permettent de définir des sous-groupes de trajectoires homogènes (Section 2.3). Nous nous sommes intéressés aux trajectoires modélisées par deux pentes et nous avons cherché à définir des sous-groupes homogènes selon le changement de pente, induit par un traitement. Nous avons donc développé une nouvelle méthode de clustering pour données longitudinales à partir de ce changement de pente, utilisant l'entièreté de la trajectoire pré- et post-traitement. De plus, nous souhaitions que la méthode développée soit robuste dans le cas de petits effectifs et de sous-groupes déséquilibrés.

3.1 Article « CLEB: a new method for treatment efficacy clustering in longitudinal data »

Problématique

Cette méthode a été développée dans le cadre de l'étude de l'efficacité des greffes dans la maladie de Huntington. Le critère de jugement pour évaluer l'efficacité de la greffe sur le plan clinique est le score moteur de l'UHDRS et l'évolution des performances motrices des patients au cours du temps. Cette échelle présente une forte variabilité intra-patient sur le plan longitudinal, qui additionnée à l'hétérogénéité d'évolution inter-patients, rend difficile l'identification des patients répondeurs à la greffe. Les 45 patients greffés de l'étude MIG-HD sont suivis longitudinalement pré- et post-greffe, ce qui nous permet de modéliser

l'évolution de leur performances motrices au cours du temps ainsi que le changement de pente au moment de la greffe. Dans le cas où la greffe n'est pas à effet immédiat, notre méthode de clustering doit tenir compte du délai entre l'initiation du traitement (la greffe) et sa prise d'effet.

Hypothèses

Une diminution des performances motrices se traduit par une augmentation du score moteur de l'UHDRS. Nous supposons qu'en l'absence de traitement ce score évolue linéairement avec le temps (Figure 14.A). Bien que l'évolution du score moteur n'est pas linéaire sur toute la durée de la maladie, l'hypothèse de linéarité sur une durée plus courte, comme celle de MIG-HD, est acceptable. Si la greffe a un effet bénéfique, elle modifie l'évolution du score, soit en ralentissant sa progression, soit en stabilisant le score, soit en inversant son évolution (Figure 14.B). Si l'effet n'est pas le même chez tous les patients, on peut voir apparaître différents profils d'évolution après l'effet du traitement (Figure 14.C).

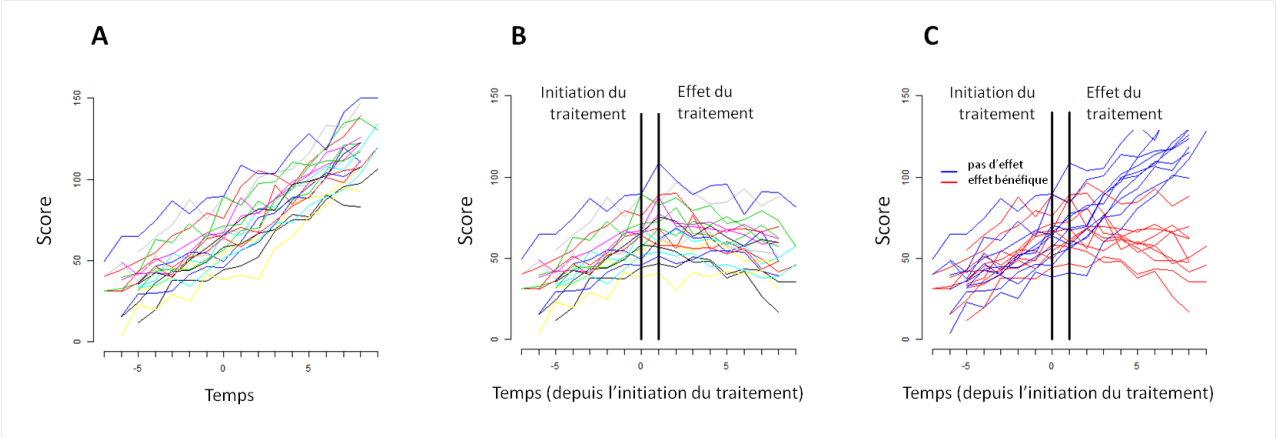


FIGURE 14 – Exemples d'évolution du score moteur de l'UHDRS avec ou sans effet du traitement (données simulées).

Les trois graphes sont des données simulées représentant l'évolution d'un score clinique (ici le score UHDRS moteur) en fonction du temps. (A) Évolution en l'absence de traitement. (B) Évolution avec effet bénéfique du traitement chez tous les patients où le score diminue après l'effet du traitement. (C) Évolution avec effet du traitement différent selon deux sous-groupes. Dans le sous-groupe avec un effet bénéfique du traitement (en rouge), le score diminue après l'effet du traitement. Dans le sous-groupe sans effet du traitement (en bleu), la pente d'évolution du score post-traitement est identique à la pente pré-traitement.

Lorsque l'effet du traitement est identique chez tous les patients, le modèle (1.9) estime correctement cet effet. Lorsque le traitement n'a pas le même effet chez tous les patients, ce modèle modélise uniquement l'effet moyen du traitement et ne permet d'estimer l'effet du traitement spécifique à chaque sous-groupe. De plus, dans ce cas, l'hypothèse de normalité des effets aléatoires n'est plus vérifiée [90]. Soit $\hat{b}_2$, l'estimation de l'effet aléatoire associé au

changement de pente dans le modèle (1.9). En présence de K sous-groupes de patients avec différents effets du traitement, nous supposons que la distribution de $\hat{b}_2$ est un mélange de K lois Gaussiennes (Figure 15).

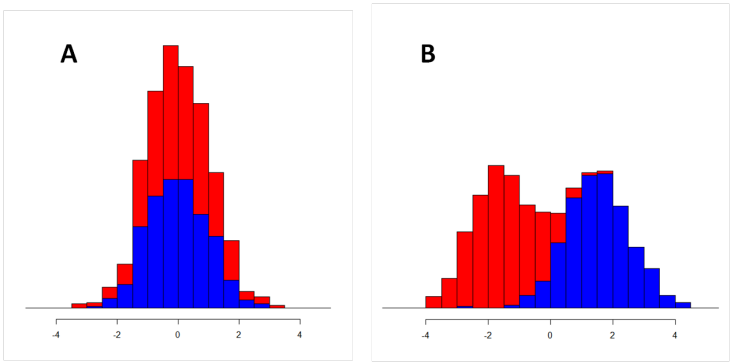


FIGURE 15 – Distribution des effets aléatoires $\hat{b}_2$ selon l'hétérogénéité de l'effet du traitement (données simulées)

(A) Les sous-groupes rouge et bleu ont le même effet du traitement. La distribution des effets aléatoires $\hat{b}_2$ suit une loi normale. (B) Les sous-groupes rouge et bleu ont un effet différent du traitement. La distribution des effets aléatoires $\hat{b}_2$ est un mélange de deux lois normales, chacune correspondant à un sous-groupe.

Le coefficient $\hat{b}_{2i}$ représente l'écart du changement de pente du patient i par rapport au changement de pente moyen de l'échantillon. Ainsi les plus grandes valeurs de $\hat{b}_{2i}$ correspondent aux patients ayant le moins de bénéfice de la greffe tandis que les plus petites valeurs de $\hat{b}_{2i}$ correspondent aux patients ayant le plus grand bénéfice de la greffe. Notre idée consiste à utiliser les valeurs de ces effets aléatoires $\hat{b}_{2i}$ pour construire les sous-groupes de patients en fonction de la réponse au traitement en les utilisant comme entrée dans les algorithmes de clustering parmi ceux décrits dans la section 2.2.

La méthode que nous proposons est en deux étapes :

1. La première étape consiste à modéliser les données longitudinales grâce au modèle à deux pentes (1.9) présenté en section 1.2 afin de récupérer les effets aléatoires de chaque patient. De cette façon, nous résumons l'information des données longitudinales (score en fonction du temps) en des données transversales (effets aléatoires).
2. La seconde étape consiste à appliquer une méthode de clustering pour données transversales sur les effets aléatoires, en particulier, sur les effets aléatoires correspondants au paramètre de changement de pente.

Afin d'évaluer les différentes stratégies (notre méthode associée à un algorithme de clustering pour données transversales), nous avons réalisé une étude de simulation.

Génération des données pour l'étude de simulation

Les différentes stratégies ont été comparées dans le cas de deux sous-groupes de patients (A et B). Les données ont été générées à partir du modèle ci-dessous :

$$y_{ij} = \begin{cases} \beta_0^{(A)} + \beta_1^{(A)} \times t_{ij} + \beta_2^{(A)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{si} \quad i \in A \\ \beta_0^{(B)} + \beta_1^{(B)} \times t_{ij} + \beta_2^{(B)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{si} \quad i \in B \end{cases} \tag{3.1}$$

Pour chaque patient, nous avons simulé 3 à 5 mesures avant et après initiation du traitement. Afin d'introduire une variabilité de délai entre deux visites j et j+1, nous avons simulé les temps de mesure tel que $t_{ij+1} - t_{ij} \sim \mathcal{N}(1,\sigma_d^2)$. Nous supposons que l'effet du traitement peut se produire à des temps variables tel que $\tau_i \sim \mathcal{N}(1/12,\sigma_\tau^2)$. Les paramètres β du modèle ont été générés pour $\ell \in \{0,1,2\}$ et $x \in \{A,B\}$ par $\beta_{\ell i}^{(x)} \sim \mathcal{N}(\mu_\ell^{(x)},\sigma_\ell^{2(x)})$. Enfin la variabilité intra-patient a été générée par $\varepsilon_{ij} \sim \mathcal{N}(0,\sigma_\varepsilon^2)$.

Nous avons fait fluctuer les valeurs de chaque paramètre individuellement pour évaluer leur impact sur le pourcentage de patients correctement classés par la méthode.

Discussion

L'article « *Clustering of Longitudinal data by using an Extended Baseline (CLEB): a new method for treatment efficacy clustering in longitudinal data* » propose une nouvelle méthode pour identifier des sous-groupes de patients selon leur réponse à un traitement dans le cas de données longitudinales. Cet algorithme intègre un algorithme classique parmi ceux présentés dans la section 2.2. Nous avons montré que notre méthode (CLEB) associée à un algorithme de clustering non paramétrique basé sur des modèles de mélanges finis est la meilleure stratégie pour classer correctement les patients dans les différents sous-groupes, même en cas de fortes variabilités intra- et inter-patients. Nos résultats montrent que cet algorithme est performant, y compris avec un fort déséquilibre entre répondeurs et non répondeurs et un petit effectif de patients. En effet, nous avons réalisé nos simulations avec un effectif total de 50 patients. Nous utilisons un algorithme de clustering plutôt que de classer les patients en fonction du signe de l'effet aléatoire car cela ne peut pas être utilisé lorsque plusieurs effets aléatoires peuvent apporter une information importante ni lorsque les sous-groupes sont de tailles différentes.

Cette méthode a été développée pour la recherche de patients « répondeurs » à la greffe dans l'essai MIG-HD. Les données peuvent provenir d'une étude de cohorte ou du bras « traitement » d'un essai clinique si celui-ci propose un suivi des patients avant l'initiation du traitement, ce qui est par exemple le cas pour le groupe contrôle d'un plan d'expérience

« *delayed-start* ». Cette méthode ne constitue pas une analyse confirmatoire, même si elle est appliquée sur des données issues d'un essai clinique. Il s'agit d'une méthode d'analyse exploratoire qui permet de générer des hypothèses sur l'explication de l'hétérogénéité de l'effet d'un traitement. Ces hypothèses peuvent ensuite être confirmées par une analyse confirmatoire dans le cadre d'un essai clinique.

Notre méthode peut être étendue pour s'adapter à des hypothèses et des données différentes :

- **Si le délai entre l'initiation du traitement et l'apparition de son effet n'est pas connu**

Nous supposons que le délai entre l'initiation du traitement et l'apparition de son effet est connu et identique pour tous les patients. Il s'agit de l'un des paramètres en entrée de notre modèle. Il serait possible d'étendre notre modèle au cas où nous n'avons pas d'*a priori* sur la valeur du délai entre l'initiation du traitement et l'apparition de son effet. Par exemple, des modèles non linéaires à base de splines permettraient de définir un « nœud » correspondant au changement de pente [91].

- **Effet du traitements à court terme et à long terme**

En modélisant les données avec deux pentes, nous avons fait l'hypothèse d'un impact constant et durable du traitement sur l'évolution du score. Cette hypothèse est forte et même si elle peut correspondre à la réalité d'un essai clinique sur une durée courte, elle ne correspond pas toujours à la réalité observée à plus long terme. En effet, un traitement peut être bénéfique, en ralentissant la progression de la maladie, durant une certaine période, puis ne plus avoir d'effet (Figure 16) ou un effet moins important voire délétère. Afin de tenir compte d'un effet du traitement à court et long terme, nous pouvons étendre notre méthode en utilisant un modèle à trois pentes. La première pente correspondra à l'évolution sans traitement, la seconde à l'évolution à court terme et la troisième à l'évolution à long terme. Une étude de simulation montrerait s'il faut tenir compte des effets aléatoires associés à la seconde et/ou à la troisième pente selon que l'on s'intéresse à l'effet du traitement à court ou à long terme ou aux deux simultanément.

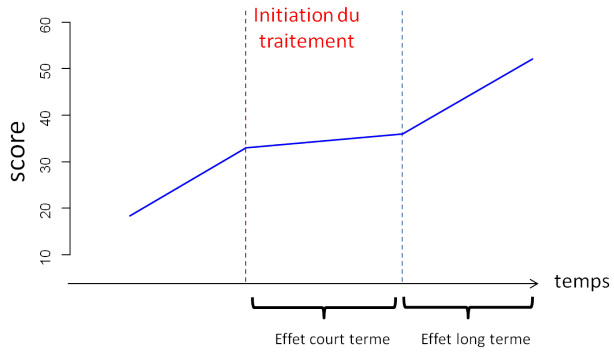


FIGURE 16 – Représentation schématique d'une évolution post-traitement à court et à long terme

Dans cet exemple, le traitement a un effet bénéfique à court terme avec une pente de progression plus faible suite à l'initiation du traitement, puis l'évolution naturelle de la maladie reprend.

- **Extension aux données d'événements récurrents**

Notre méthode a été développée pour des données quantitatives longitudinales. Elle peut être adaptée à d'autres types de données. Par exemple, nous l'avons étendu aux données d'événements récurrents. Dans ce cas, l'effet du traitement est mesuré par la différence des délais d'apparition des événements pré- et post-traitement. La première étape de notre méthode, consistant à modéliser les données, est réalisée grâce à une adaptation du modèle de Cox pour événements récurrents et intégrant des paramètres aléatoires. Le détail de notre méthode étendue aux événements récurrents ainsi qu'une étude de simulation sont présentés en Annexe B. Nous avons implémenté notre méthode CLEB et son extension aux événements récurrents (*Clustering of Recurrent Events using Mixed Effects*, CREME) dans un package R qui sera accessible sur le site du CRAN. Pour des données binaires ou catégorielles mesurées longitudinalement, notre méthode peut s'adapter en remplaçant le modèle linéaire mixte par un modèle linéaire mixte généralisé [92].

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–21

DOI: 10.1177/0962280215621591
smm.sagepub.com
SAGE

Clustering of longitudinal data by using an extended baseline: A new method for treatment efficacy clustering in longitudinal data

Catherine Schramm,[1,2,3,4] Céline Vial,[5] Anne-Catherine Bachoud-Lévi[2,4,6] and Sandrine Katsahian[1,7]

Abstract

Heterogeneity in treatment efficacy is a major concern in clinical trials. Clustering may help to identify the treatment responders and the non-responders. In the context of longitudinal cluster analyses, sample size and variability of the times of measurements are the main issues with the current methods. Here, we propose a new two-step method for the Clustering of Longitudinal data by using an Extended Baseline. The first step relies on a piecewise linear mixed model for repeated measurements with a treatment-time interaction. The second step clusters the random predictions and considers several parametric (model-based) and non-parametric (partitioning, ascendant hierarchical clustering) algorithms. A simulation study compares all options of the clustering of longitudinal data by using an extended baseline method with the latent-class mixed model. The clustering of longitudinal data by using an extended baseline method with the two model-based algorithms was the more robust model. The clustering of longitudinal data by using an extended baseline method with all the non-parametric algorithms failed when there were unequal variances of treatment effect between clusters or when the subgroups had unbalanced sample sizes. The latent-class mixed model failed when the between-patients slope variability is high. Two real data sets on neurodegenerative disease and on obesity illustrate the clustering of longitudinal data by using an extended baseline method and show how clustering may help to identify the marker(s) of the treatment response. The application of the clustering of longitudinal data by using an extended baseline method in exploratory analysis as the first stage before setting up stratified designs can provide a better estimation of treatment effect in future clinical trials.

[1]INSERM UMRS1138, Centre de Recherche des Cordeliers, E22, Université Paris Descartes, Université Pierre et Marie Curie, Paris, France
[2]INSERM U955 E01, Neuropsychologie interventionnelle Laboratory IMRB, Créteil, France
[3]Université Pierre et Marie Curie, Paris 6, Paris, France
[4]École Normale Supérieure, Institut d'Études de la Cognition, Paris, France
[5]Université de Lyon, CNRS UMR 5208, Polytech Lyon-Université de Lyon 1, Institut Camille Jordan, Villeurbanne, France
[6]Assistance Publique-Hôpitaux de Paris, National Reference Center for Huntington's Disease Henri Mondor Hospital, Créteil, France
[7]Assistance Publique-Hôpitaux de Paris, Service d'informatique et statistiques, Georges Pompidou European Hospital, Paris, France

Corresponding author:
Catherine Schramm, Centre de Recherche des Cordeliers, Escalier D, 1er étage 15 rue de l'école de médecine 75006 Paris, France.
Email: cath.schramm@gmail.com

Keywords
Clustering, longitudinal data, personalized medicine, treatment effect, Huntington's disease, obesity

1 Introduction

Heterogeneity in treatment efficacy is one of the biggest concerns in personalized medicine. However, clustering may help to identify the treatment responders and the non-responders. Clustering is an unsupervised learning method that allows a hidden structure to be found in unlabelled data. It relies on an algorithm to minimize within-cluster variability (internal cohesion) and maximize between-cluster variability (external isolation).[1] Here, we explore the specific context of a rare and progressive disease with a small sample size and a treatment effect that is measured longitudinally. Thus, the information provided by the longitudinal data is not restricted to a single value but must consider the entire trajectory of a continuous score. More precisely, we focus on the change of slope after the treatment initiation. For this longitudinal cluster analysis, the data may be obtained from cohort studies or from clinical trials (treatment arm) with the treatment initiation during the follow-up. In both cases, repeated measurements of the patients' scores are recorded before and after the initiation of treatment.

The current parametric and non-parametric methods for longitudinal cluster analysis are being increasingly used in medical research.[2–6] Parametric methods relate to mixture modeling techniques, in particular through latent-class mixed models (LCMM). They assess the influence of latent growth trajectory class membership on the outcome to highlight the distinct patterns of evolution.[7] The mixed model allows the within-subject correlation and the variability of the outcome trajectory between subjects to be taken into account. Latent-class is defined by using the assumption of a mixture of Gaussian distributions for the random effects.[8] Clusters and model parameters are estimated simultaneously. The main advantage of these parametric methods is that the usual statistical tests and inferences can be performed; however, if they are to be efficient these methods often require a large sample of patients, which might not be the case for rare diseases or innovative therapies like cell transplant or gene therapy. Non-parametric methods relate to classical algorithmic approaches such as K-Means for Longitudinal data (KML).[9] Such algorithms consider the distance between patients' score rather than the shape of the evolution which does not address the initial problem of the change in the slope due to treatment effect. Furthermore, they need a constant measurement delay between patients, which is not a reasonable assumption in cohort studies. The limits of these methods suggest the need for a new method to cluster longitudinal data according to treatment effect when there is both a small sample and variability in the times of measurement. We propose a Clustering of Longitudinal data with an Extended Baseline (CLEB) method. This new method comprises two steps: first, building a linear mixed model with an extended baseline and second clustering the random predictions through a model-based algorithm. However, other strategies of clustering could be planned in the second step, notably non-parametric algorithms.

The objectives of this paper are (i) to present this new method and (ii) to compare the model-based and non-parametric strategies. The CLEB method is described in section CLEB. The simulation study settings and results are presented in, respectively, sections **The simulation study procedure** and **Results of the simulation study**. The different strategies of the CLEB method are evaluated and compared to the LCMM algorithm. The CLEB method is illustrated in section

Applications with a real data set of Huntington's disease patients and a real data set of women suffering from obesity. The results are discussed in section Discussion.

2 CLEB

The CLEB method clusters patients according to treatment effect. The two steps of this method are described in this section.

Consider data from $i = 1, ., N$ patients in a longitudinal study assessed n_i times with y_{ij} the jth outcome measure of patient i at time t_{ij}. Patients initiate their treatment during the follow-up and the times are realigned such that 0 corresponds to the time of treatment initiation. Thus, there are negative times ($t_{ij} < 0$) for measurements before the treatment initiation. The lag $\tau \geq 0$ between treatment initiation and treatment effect is used to define two phases: the baseline phase for $t_{ij} < \tau$ and the treatment effect phase for $t_{ij} \geq \tau$.

2.1 Step 1: The mixed model with extended baseline for treatment-time interaction

Figure 1 represents the CLEB algorithm.

The basic assumption of the method is that the individual's responses vary linearly according to phase. The polynomial model with extended baseline[10] could be adapted to the data as follows

$$y_{ij} = a_{0i} + a_{1i} \times t_{ij} + a_{2i} \times (t_{ij} - \tau) \times 1(t_{ij} \geq \tau) + \varepsilon_{ij} \quad (1)$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

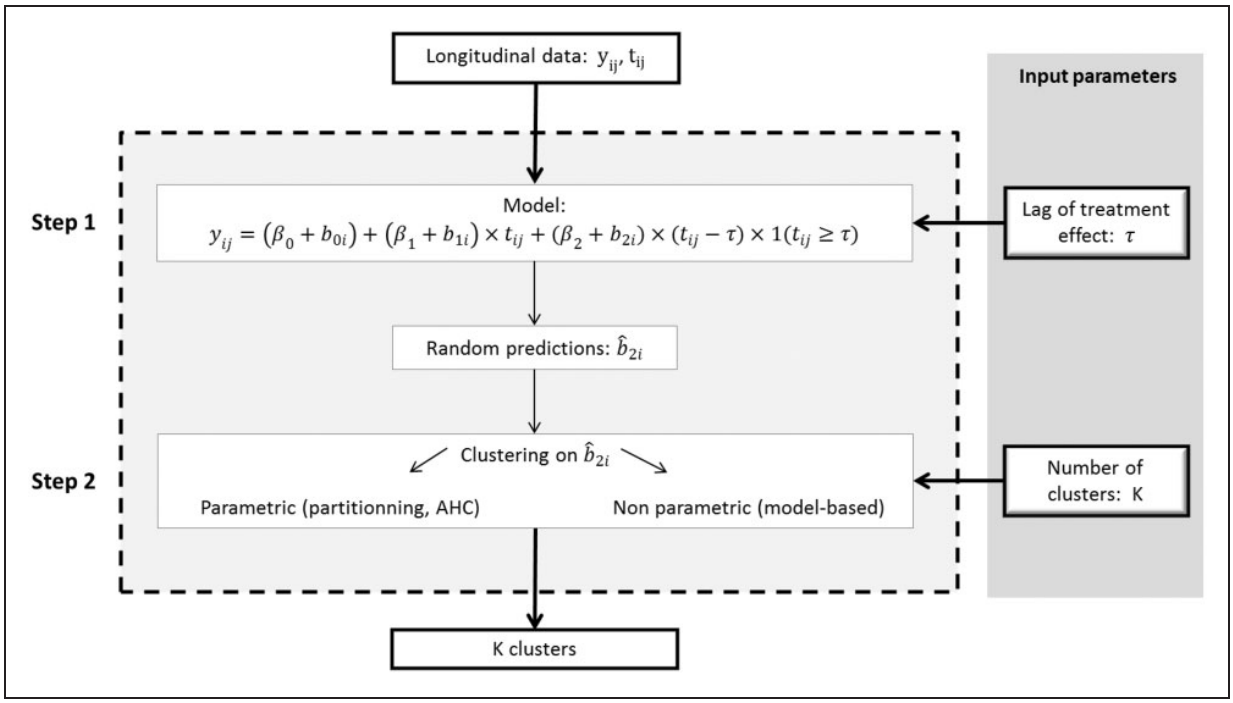


Figure 1. General architecture of the CLEB method.

The random effect model for longitudinal data[11] takes into account all the available information, allowing the model to deal with missing data and numbers and times of measurements that are not identical.[12] It takes into account both within- and between-patient variability. Assuming between-patient variability in the baseline phase (intercept and slope) and in the treatment effect phase, $\boldsymbol{a_i}$ could be decomposed as the sum of a fixed effect $\boldsymbol{\beta}$ and a random effect $\boldsymbol{b_i} \sim \mathcal{N}(0, \boldsymbol{D})$ with $\boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix}$ such that equation (1) becomes

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t_{ij} + (\beta_2 + b_{2i}) \times (t_{ij} - \tau) \times 1(t_{ij} \geq \tau) + \varepsilon_{ij} \quad (2)$$

The mean score at treatment initiation is β_0. The mean progression slope of y_{ij} during the baseline phase is β_1 and the mean progression slope of y_{ij} during the treatment effect phase is the sum $\beta_1 + \beta_2$, where β_2 is the fixed estimate associated with the treatment effect. The third component of equation (2) can be considered as a time-treatment interaction.

The estimation of parameters is made using restricted maximum likelihood, but only the predictions of the random parameters ($\hat{\boldsymbol{b_0}}$, $\hat{\boldsymbol{b_1}}$ and $\hat{\boldsymbol{b_2}}$) are collected. They will be used in the second step of the CLEB method.

2.2 Step 2: Clustering on random predictions

The distribution of $\hat{\boldsymbol{b_2}}$ is assumed to be a mixture of K Gaussian distributions, each distribution corresponding to one cluster.[8] For two clusters A and B of treatment effect $\hat{b}_{2i} \sim p\mathcal{N}(\mu_{2,A}, \sigma^2_{2,A}) + (1-p)\mathcal{N}(\mu_{2,B}, \sigma^2_{2,B})$, where p (resp. $1-p$) is the proportion of subjects in cluster A (resp. B), $\mu_{2,A}$ (resp. $\mu_{2,B}$) and $\sigma^2_{2,A}$ (resp. $\sigma^2_{2,B}$) are the mean and variance of the individual treatment effect ($\hat{b}_{2i}$) of the patients in cluster A (resp. B). The second step of the CLEB method consists of estimating the mixture of distributions using an Expectation-Maximization (EM) model-based algorithm. EM model-based algorithm and non-parametric algorithms are presented in the following subsections.

2.2.1 EM model-based algorithm

Let us now briefly describe the model-based algorithm to cluster patients using $\hat{\boldsymbol{b_2}}$. This parametric model supposes a Gaussian distribution of $\hat{\boldsymbol{b_2}}$ for each cluster.[13] Let f be the density function of the mixture defined by

$$f(\hat{\boldsymbol{b_2}}) = \sum_{k=1}^{K} \boldsymbol{\pi_k} \Phi(\hat{\boldsymbol{b_2}} | \mu_{2,k}, \sigma^2_{2,k}) \quad (3)$$

where $\boldsymbol{\pi_k}$ is the probability that a subject belongs to the cluster k and $\Phi(\hat{\boldsymbol{b_2}} | \mu_{2,k}, \sigma^2_{2,k})$ is the density function from the distribution $\mathcal{N}(\mu_{2,k}, \sigma^2_{2,k})$. The method uses the maximization of likelihood in the EM algorithm to estimate $\mu_{2,k}$ and $\sigma^2_{2,k}$ for each cluster k and $\pi_{k,i}$ for each cluster k and each patient i.[14] Two parameterizations could be considered:

- E parameterization: equal variance between clusters
- V parameterization: variable variances between clusters

Then $\pi_{k,i}$ is used to classify each patient in the cluster with the higher probability such that $i \in k'$ if $\pi_{k',i} = \max_k \pi_{k,i}$.

This algorithm could be extended to the case of clustering on the multivariate component $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_1)$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$. In this case, the density function g of the multivariate mixture is defined by

$$g(\boldsymbol{\theta}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \Phi(\boldsymbol{\theta}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

where $\Phi(\boldsymbol{\theta}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the density function from the multivariate normal distribution $\mathcal{N}_{dim(\boldsymbol{\theta})}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\mu}_k$ the vector of means and $\boldsymbol{\Sigma}_k$ the variance–covariance matrix.

In the multivariate case, ten parameterizations may be considered. They concern the variance–covariance matrix structure defined according to three geometric parameters: volume (equal: E or variable: V), shape (equal: E, variable: V, identity: I) and orientation (equal: E, variable: V, identity: I) between clusters.[15]

2.2.2 *Non-parametric alternatives (k-means, k-medoids, agglomerative hierarchical clustering)*

The k-means[16,17] and the k-medoids[18] are iterative algorithms partitioning the data space into Voronoi cells. They attribute data points to clusters by minimizing the distance between each point and the mean of the cluster (in k-means) or the most central value (in k-medoids).

Agglomerative Hierarchical Clustering (AHC) procedures[19] agglomerate the data from N single-member clusters into one cluster containing all data points and stop when the expected number of clusters is reached. The AHC-single linkage, -complete linkage and -average linkage define the distance between clusters as, respectively, the minimal, maximal and average distance between the data points of each cluster.

Several definitions for calculating the distances between two data points exist. Those considered in the simulation study are defined below. Let $\boldsymbol{\theta}$ be the vector of random predictions such that $\boldsymbol{\theta} = \hat{\boldsymbol{b}}_2$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_1)$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$. The Euclidean, Canberra, Manhattan, Maximum, Pearson, and Correlation distances for two patients i and j are defined as

$$d_{\text{Euclidean}}(i,j) = \sqrt{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} (\hat{b}_{\ell,i} - \hat{b}_{\ell,j})^2}$$

$$d_{\text{Canberra}}(i,j) = \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \frac{|\hat{b}_{\ell,i} - \hat{b}_{\ell,j}|}{|\hat{b}_{\ell,i}| + |\hat{b}_{\ell,j}|}$$

$$d_{\text{Manhattan}}(i,j) = \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} |\hat{b}_{\ell,i} - \hat{b}_{\ell,j}|$$

$$d_{\text{Maximum}}(i,j) = \max_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} |\hat{b}_{\ell,i} - \hat{b}_{\ell,j}|$$

$$d_{\text{Pearson}}(i,j) = \frac{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}_{\ell,i} \hat{b}_{\ell,j}}{\sqrt{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}_{\ell,i}^2 \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}_{\ell,j}^2}}$$

$$d_{\text{Correlation}}(i,j) = \frac{Cov(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}{\sqrt{Var(\boldsymbol{\theta}_i) Var(\boldsymbol{\theta}_j)}}$$

Note that for $\boldsymbol{\theta} = \hat{\boldsymbol{b}}_2$, the Pearson distance is always equal to one, the correlation distance is not defined and the Manhattan and Maximum distances are equal to the Euclidean distance. Thus only the Euclidean and Canberra distances will be used in the univariate case.

3 The simulation study procedure

Several scenarios with two subgroups of patients (N_A subjects with a beneficial treatment effect in group A and N_B subjects with a detrimental treatment effect in group B) were considered. The notation is similar to that in section CLEB. Each patient i has at least one visit before the treatment initiation, one visit at treatment initiation and one visit after treatment initiation. The period between visits ($\boldsymbol{d}$) is around one year: $d_{ij} \sim \mathcal{N}(1, \sigma_d^2)$ such that $t_{ij} - t_{ij-1} = d_{ij}$. The outcome for each visit is generated by

$$y_{ij} = \begin{cases} \beta_0^{(A)} + \beta_1^{(A)} \times t_{ij} + \beta_2^{(A)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{if} \quad i \in A \\ \beta_0^{(B)} + \beta_1^{(B)} \times t_{ij} + \beta_2^{(B)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{if} \quad i \in B \end{cases} \tag{5}$$

with $\tau_i \sim \mathcal{N}(1/12, \sigma_\tau^2)$ meaning that the treatment effect is supposed to appear at one month. For $\ell \in \{0, 1, 2\}$ and $x \in \{A, B\}$, $\beta_{\ell i}^{(x)} \sim \mathcal{N}(\mu_\ell^{(x)}, \sigma_\ell^{2(x)})$. The strength of treatment effect is $|\mu_2^{(A)} - \mu_2^{(B)}|$, and the within-treatment variation in group A (respectively B) is $\sigma_2^{(A)}$ (respectively $\sigma_2^{(B)}$). The within-patient variability is generated by $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The simple case scenario is parameterized with 25 patients per group ($N_A = N_B = 25$), each having 3 to 5 visits before treatment initiation and 3 to 5 visits after treatment initiation, null variance on the period between visits and the lag in treatment effect ($\sigma_d = \sigma_\tau = 0$) and low within-patient variability ($\sigma_\varepsilon = 3$). The mean of the effects in group A ($\mu_0^{(A)}, \mu_1^{(A)}, \mu_2^{(A)}$) is the vector (45, 5, $-$ 5), whereas the mean of the effects in group B ($\mu_0^{(B)}, \mu_1^{(B)}, \mu_2^{(B)}$) is the vector (45, 5, 5). Finally, the between-patient variabilities in groups A ($\sigma_0^{(A)}, \sigma_1^{(A)}, \sigma_2^{(A)}$) and B ($\sigma_0^{(B)}, \sigma_1^{(B)}, \sigma_2^{(B)}$) are both initialized by the vector (5, 1, 1).

The baseline mean parameters were initialized according to the pre-randomization period in a longitudinal clinical trial assessing graft in Huntington's disease (NCT00190450), for which the CLEB method was developed. However, this clinical trial has not yet been published, and the data could not be used as an illustrated example in section Applications.

The simulation study compares the CLEB method, with the different strategies, to the LCMM method. The LCMM method assumes that each cluster, also called the latent-class, is characterized by a specific trajectory modelled by a specific linear mixed model. Both the latent-class membership and the trajectory are explained using covariates. Here, the parameterization of the LCMM method for the trajectory was performed in the same way as in equation (2). Only the third term, that corresponding to the time-treatment interaction was used in the parameterization of the latent-class membership. Likelihood maximization and the EM algorithm were used to estimate the class membership probability and the model parameters simultaneously.

All results are expressed as the mean of the percentage of correctly classified patients among the 1000 databases generated for each scenario.

The simulations and computations were performed using the R software.[20] The CLEB method was performed using the `nlme`[21] package for a mixed model for step 1, the `mclust`[22] package for

model-based clustering for step 2, and the `amap`[23] and `cluster`[24] packages for non-parametric algorithms for step 2. The LCMM method was performed using the `lcmm`[25] package.

4 Results of the simulation study

Only the most efficient strategies are presented. Thus, for the CLEB method, the AHC strategies are not reported, and k-means were preferred to k-medoids. For the same reason, except for the specific scenarios in which the treatment effect was influenced by the baseline parameters, only univariate strategies are presented.

Figure 2 displays the percentage of correctly classified patients according to the strength of treatment effect and sample size. When there was a great difference between the simulated treatment effects for the two groups, all strategies allocated almost 100% of patients to the correct cluster. In contrast, when there was no difference between the two groups, all strategies randomly allocated patients to each cluster (50% of correctly classified patients). The CLEB method gave better results than the LCMM regardless of the strategy that was used. The sample size had a greater impact for the CLEB method with the model-based strategies, with a better classification for a larger sample size.

Figure 3 displays the percentage of correctly classified patients according to the natural disease progression variability (variability of slope during the baseline phase: σ_1) and the within-treatment variability (variability of slope change: σ_2). Whatever the strategy, the CLEB method was not affected by the natural disease progression variability, whereas the performances of the LCMM method were worse when σ_1 was greater (Figure 3(a)). Indeed, in LCMM, the subgroup identification and the estimation of the parameters were made simultaneously, leading to a greater influence of the baseline slope on the subgroup definition. The more σ_2 increased, the worse the performances of all methods (Figure 3(b) and (c)). As expected, in CLEB with the EM model-based algorithm, the V parameterization (variable variances between clusters) suffered less of an impact from high variance than the E parameterization (equal variances between clusters) by high variance when the variances were unequal, whereas the V parameterization was equivalent to the E parameterization in the case of equal variances.

Figure 4(a) displays the percentage of correctly classified patients according to the number of subjects in each subgroup. All methods had good performance when the groups were balanced. However, the CLEB method with k-means strategy and Canberra distance did not perform well in the case of unbalanced groups. The LCMM method and the CLEB method with k-means and Euclidean distance were only unable to perform well in an extremely unbalanced case ($N_A = 2$). The CLEB method with EM model-based algorithms was the only model that was not affected at all by unbalanced groups. In the case of $N_A = 2$, the CLEB method with EM model-based algorithms found a small subgroup of responders (Figure 4(b)). All methods have good sensitivity, but the CLEB method with the two EM model-based algorithms has the better specificity.

Figure 5 displays the percentage of correctly classified patients according to the number of time points. For low variability of natural disease progression, an increase in the number of time points after treatment initiation improved the performance of all methods, whereas an increase in the number of time points before treatment initiation did not have an impact on their performances. For high variability of natural disease progression, an increase in the number of time points after and/or before treatment initiation improved the performance of the CLEB method, regardless of the associated strategy. The LCMM method, for which variability of natural disease progression had a real impact (Figure 3), did not have an improved performance, regardless of the number of time points. These results suggest that clusters of treatment effects could be identified using only the treatment effect phase in the case of a homogeneous natural disease progression. However, in the case of a heterogeneous natural disease progression, the slope of the baseline phase

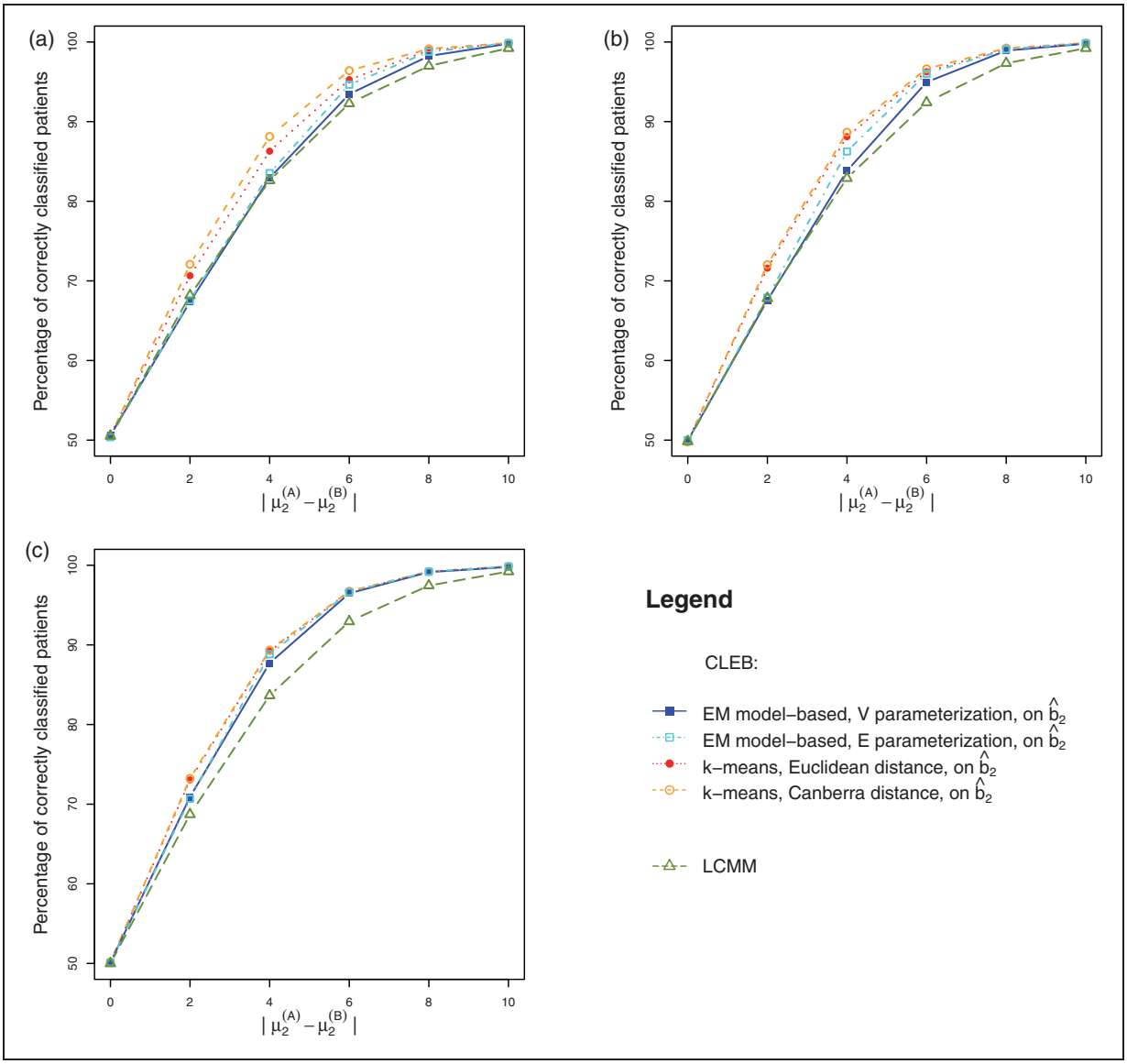


Figure 2. Impact of the strength of treatment effect and the sample size on the percentage of correctly classified patients. (a) $N_A = N_B = 10$; (b) $N_A = N_B = 25$ and (c) $N_A = N_B = 100$.

is necessary to evaluate the treatment effect on the slope change between the baseline and the treatment effect phases.

Figure 6 displays the percentage of correctly classified patients when the baseline characteristics differ in the two clusters. For parametric algorithms, only the better univariate and multivariate strategies are presented; these correspond to, respectively, the V and VVI parameterizations. It should be noted that V corresponds to a hypothesis of variable variance and VVI to a hypothesis of variable variance according to clusters and according to random terms without correlation between clusters. For the non-parametric algorithms, only the k-means with Euclidean distance is presented. Other non-parametric algorithms show similar results. The multivariate strategies improve when the correlation between baseline and treatment effect increases, because $\hat{b}_0$ and $\hat{b}_1$

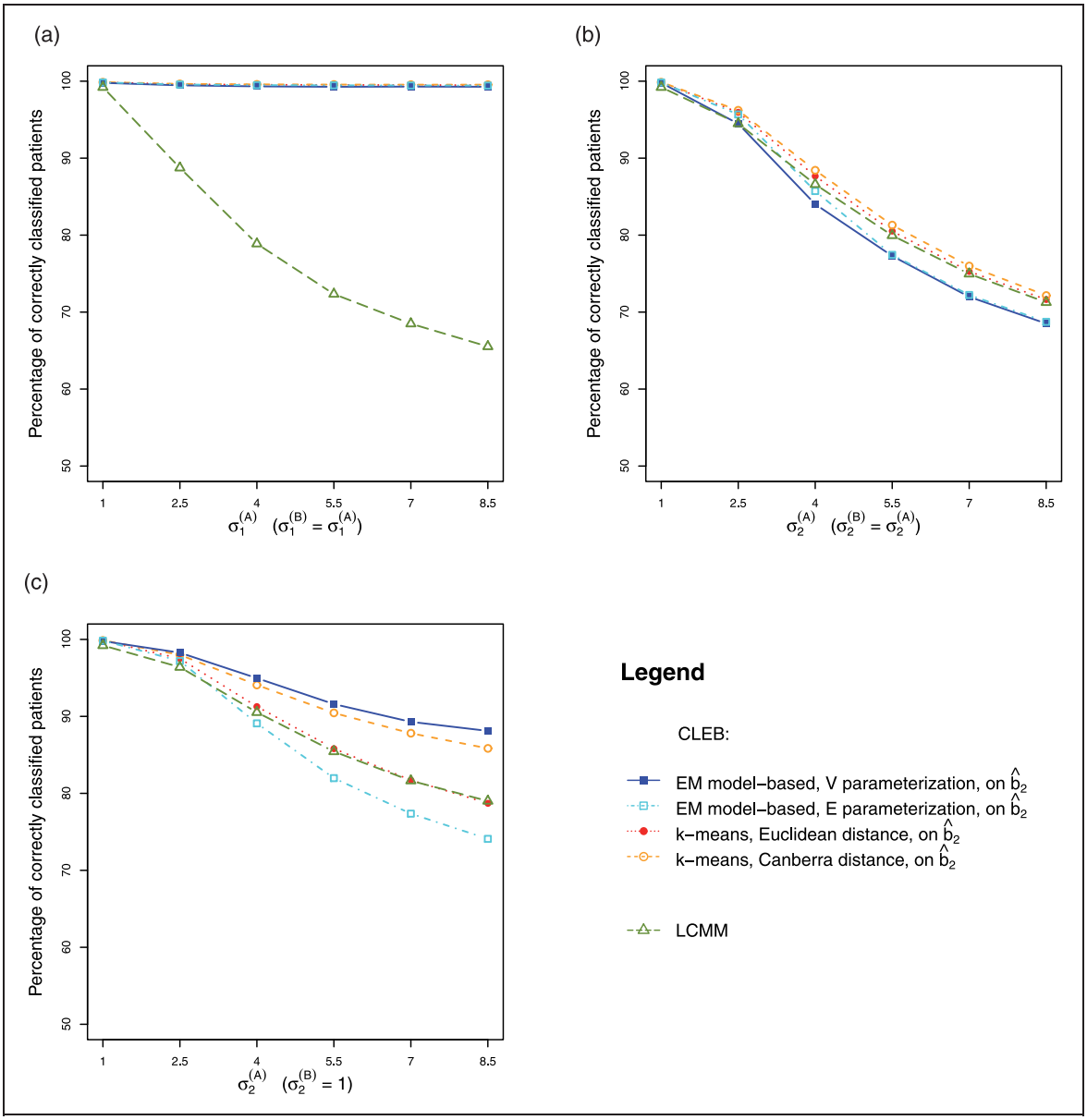


Figure 3. Impact of heterogeneity of the natural disease progression and the within-treatment variation on the percentage of correctly classified patients. (a) Heterogeneity of natural disease progression; (b) Within-treatment variation in two groups and (c) Within-treatment variartion in responders.

also capture some useful information for clustering. However, if variability of the baseline parameters is high, the use of these parameters will lead to a less robust classification than the univariate strategies (Figure 6(c) and (f)). The univariate strategy may also be improved by an increase in the difference between the natural slope of the disease in the two subgroups (Figure 6(b) and (e)). Indeed, the difference in natural evolution increases the difference in the slope after treatment initiation, leading to a better classification.

Figure 7 displays the percentage of correctly classified patients in the case of missing data. When the methods were applied using the whole data set, patients with missing data were allocated

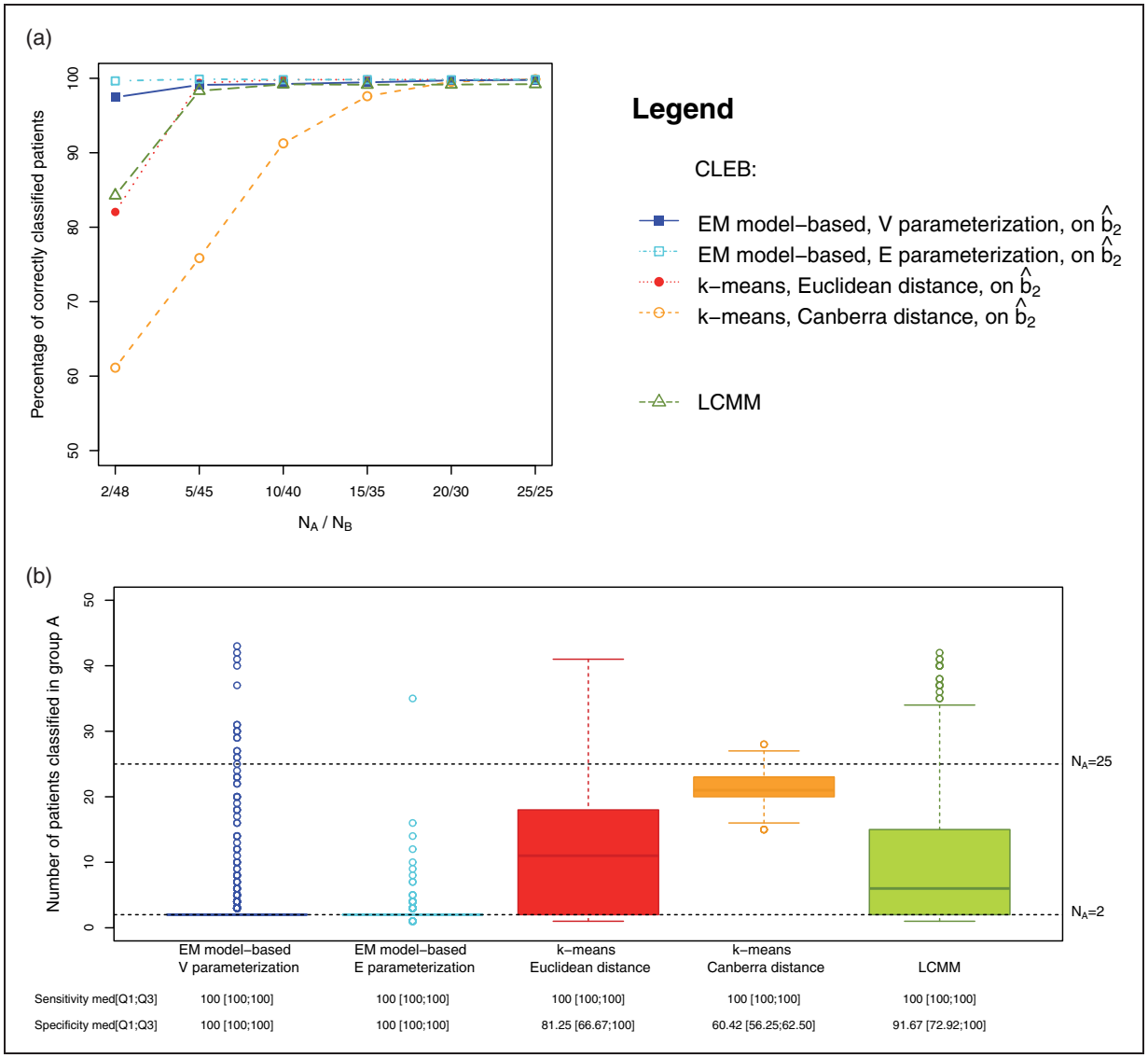


Figure 4. The case of unbalanced clusters. (a) Impact on classification and (b) Specific case of $N_A = 2$ and $N_B = 48$.

randomly into clusters (Figure 7(c)), whereas almost 100% of the patients without missing data were allocated to the correct cluster (Figure 7(d)). Only the CLEB method with the EM model-based algorithm and V parameterization was slightly affected by a high rate of missing data. However, applying the method only to subjects without missing data (the complete case study) led to the best results (Figure 7(b)).

The simulation study showed that the CLEB method performs better than LCMM when there is high slope variability. In the CLEB method, univariate strategies were preferred to multivariate strategies. The k-means with Canberra distance was hugely affected by unbalanced groups, to the extent that it was not a reliable strategy. The EM model-based algorithm with V parameterization (the hypothesis of variable variance between clusters) must be the preferred strategy, but the CLEB

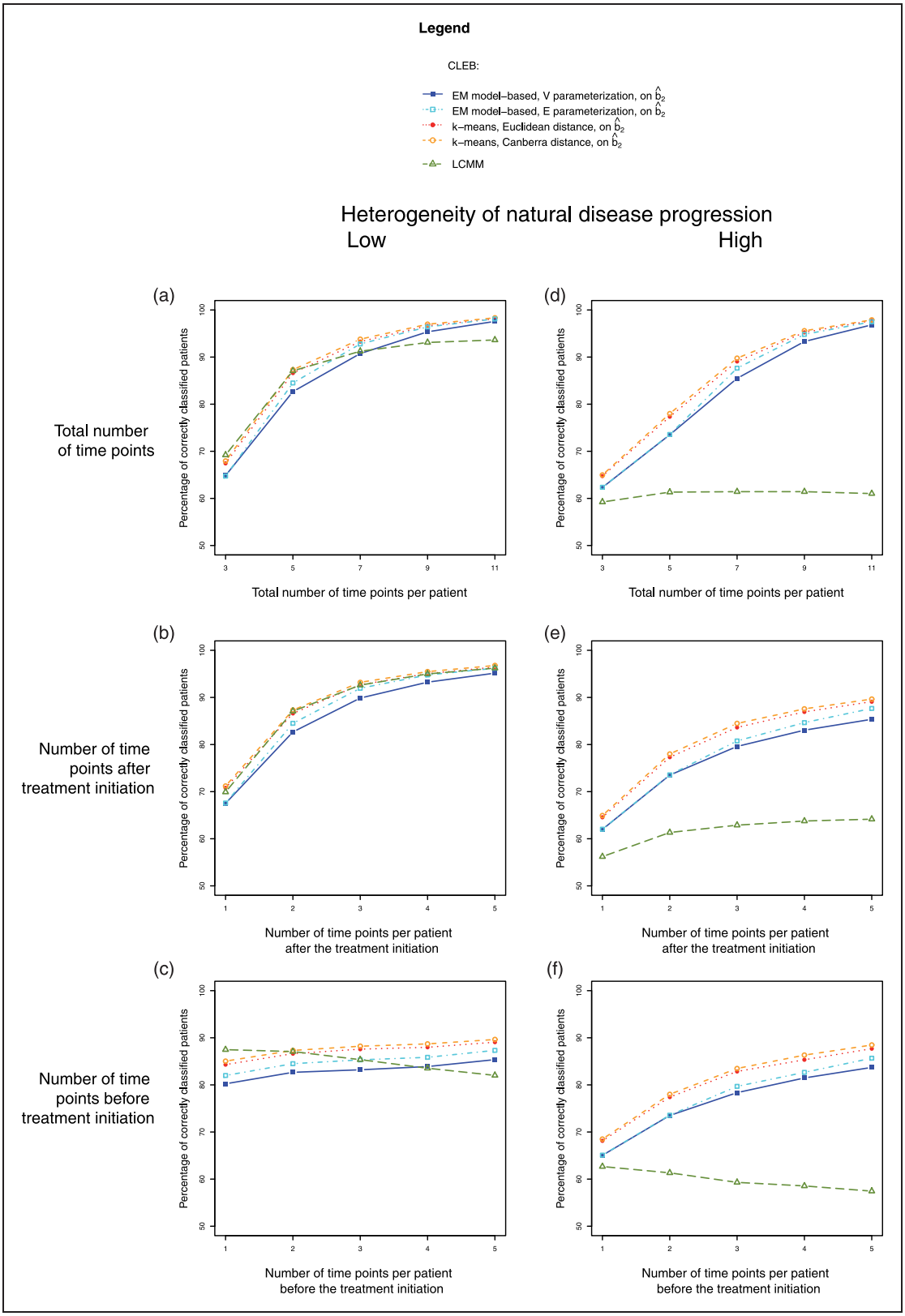


Figure 5. Impact of the number of time points on the percentage of correctly classified patients.

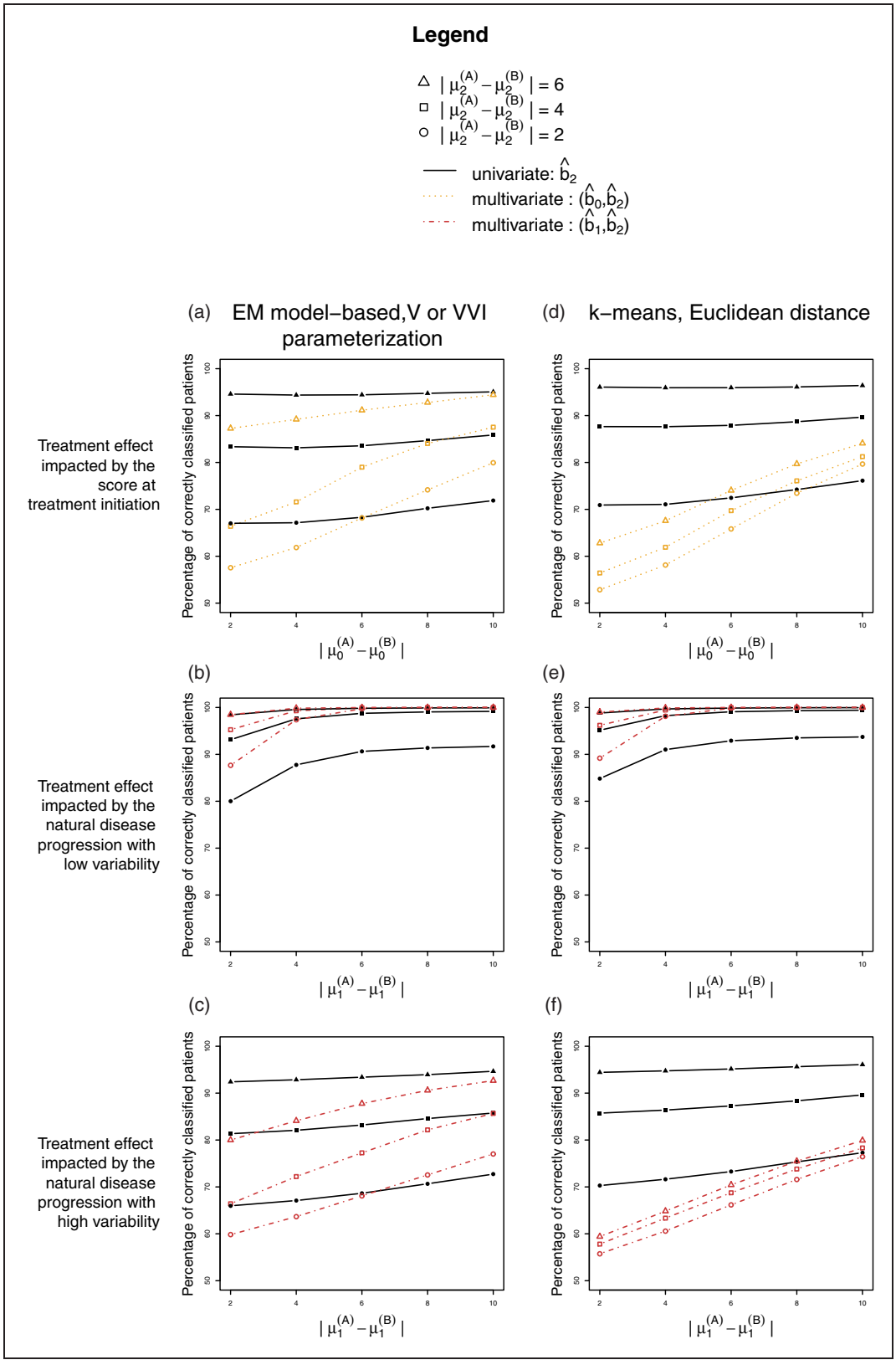


Figure 6. The case of a treatment effect on which baseline or evolution before treatment initiation has an impact.

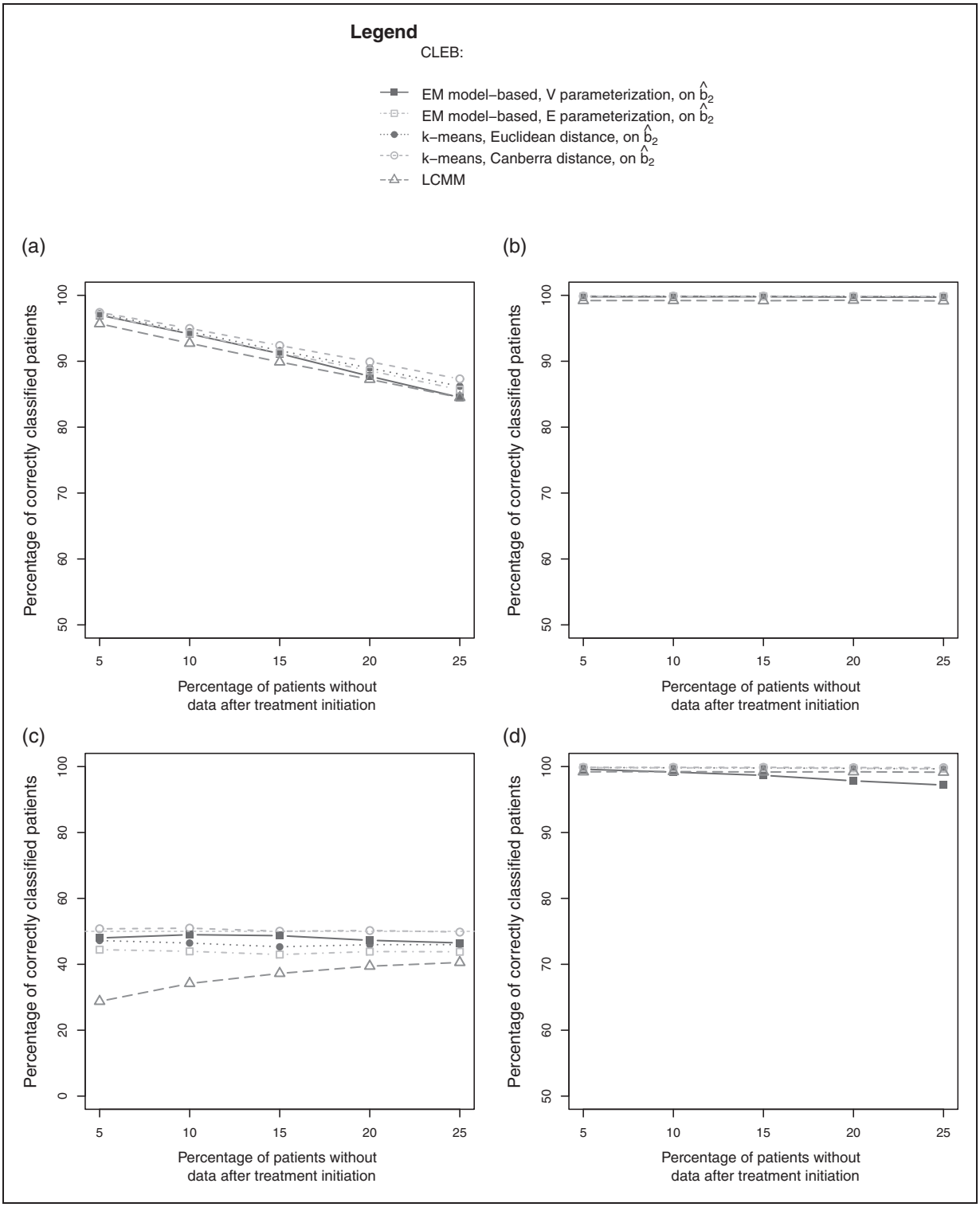


Figure 7. Impact of missing data on the classification. (a) Analysis with all patients (patients with and without missing data); (b) Analysis with complete cases (only patients without missing data); (c) Results for patients with missing data in analysis with all patients and (d) Results for patients without missing data in analysis with all patients.

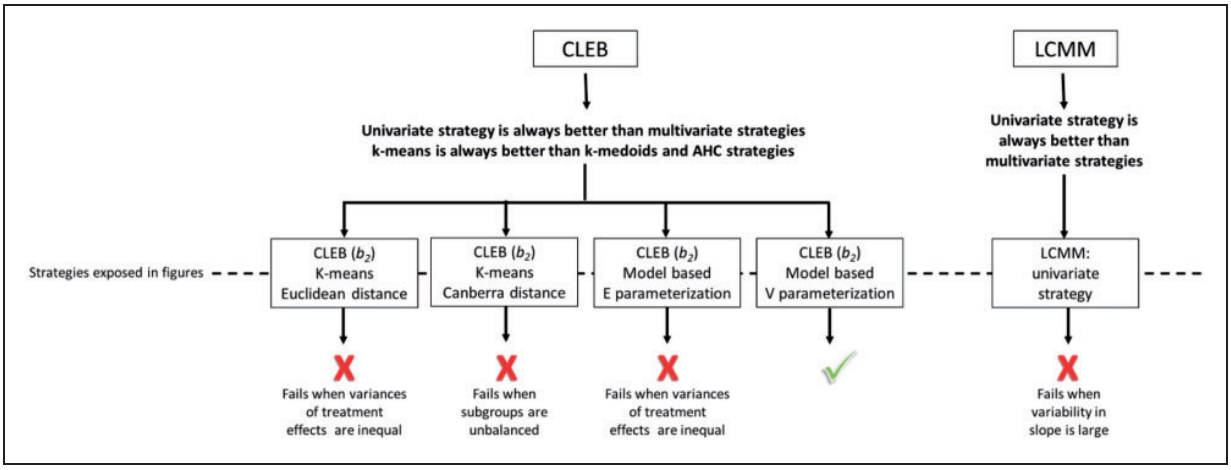


Figure 8. Results of the simulation study.

method with k-means algorithm and Euclidean distance could be more robust when there is a small sample size and low variance. Figure 8 sums up all the strategies considered and shows how the simulation study reached this conclusion.

5 Applications

5.1 The impact of neuroleptics on the evolution of Huntington's disease

Huntington's disease is a rare and inherited neurodegenerative disorder caused by an expansion of a CAG (Cytosine-Adenine-Guanine) triplet repeat on the huntingtin gene on chromosome 4. It is characterized by choreiform movements, progressive dementia and psychiatric manifestations.[26] There is currently no cure and all available treatments are symptomatic, i.e. they treat the symptoms but not the underlying disease. For example, AntiPsychotics and Related drugs (APRs) are commonly used for the treatment of chorea. Here, we evaluate the response to treatment with APRs.

We searched for responders to the treatment based on the evolution of the Functional Assessment Score (FAS), a clinical marker of the progress of the disease (score from 25 to 50). The treatment was supposed to start taking effect one month after the first prescription.

Data were selected from the Huntington French Speaking Network cohort between 2002 and 2010, among the patients studied by Désaméricq et al.[27] In this clustering study, only 39 patients having APRs treatment who were followed at least twice before and twice after the treatment initiation were included. They were followed for 4.98 years ($SD = 1.58$), representing between 4 and 12 visits.

We applied the CLEB method (with the model-based algorithm and V parameterization) to the 39 patients. The CLEB splitted the population into two subgroups: 15 responders and 24 non-responders to APRs treatment. We then modelled the data with these groups of treatment responses as the covariate. The model showed an evolution of FAS of 1.43 points per year ($SE = 0.22$, $P < 0.001$) during the baseline phase in the whole cohort. The difference in slope between the baseline phase and the treatment effect phase was –0.03 points per year ($SE = 0.35$, $P = 0.930$) for responders and 2.08 points per year ($SE = 0.37$, $P < 0.001$) for non-responders (Figure 9).

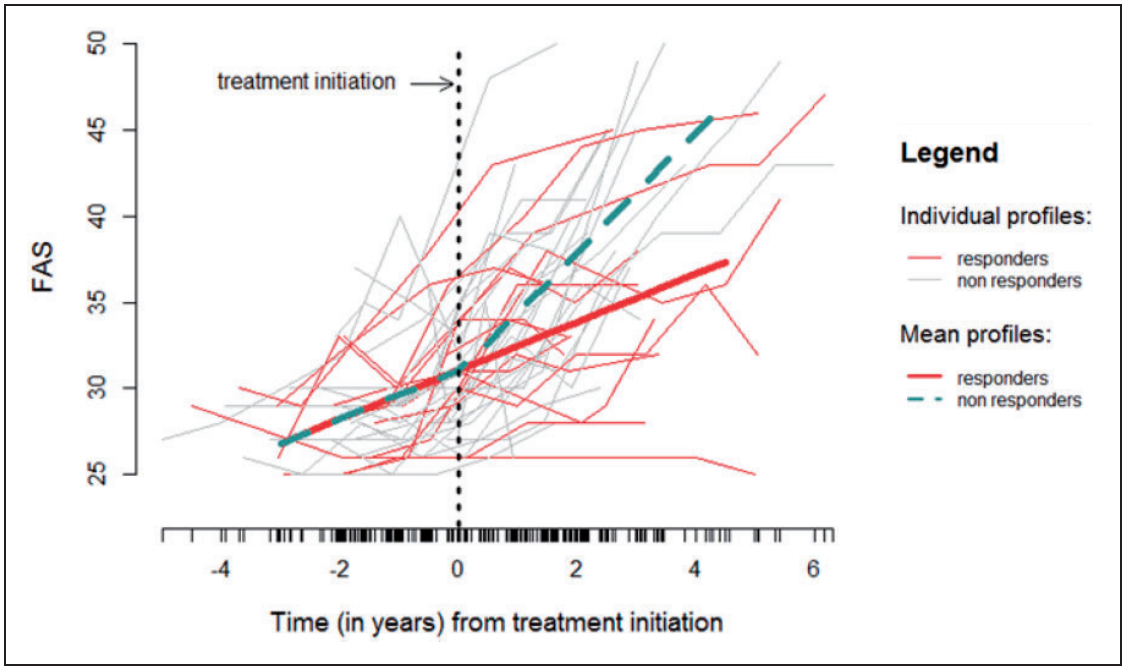


Figure 9. Evolution of FAS in the subgroups of responders and non-responders.

Table 1. Description of responders and non-responders to APRs.

	Whole cohort $N = 39$	Responders $N = 15$	Non-responders $N = 24$	p-values[b]
Age[a] (y)	50.07 (8.81)	52.31 (6.86)	48.67 (9.71)	0.260
Sex				
Male	23 (58.97%)	9 (60.00%)	14 (58.33%)	0.918
Female	16 (41.03%)	6 (40.00%)	10 (41.67%)	
Inheritance				
Paternal	19 (48.72%)	8 (53.33%)	11 (45.83%)	0.676
Maternal	17 (43.59%)	6 (40.00%)	11 (45.83%)	
Unknown	3 (7.69%)	1 (6.67%)	2 (8.33%)	
CAG	44.23 (3.06)	42.60 (2.47)	45.25 (3.00)	0.008
Age at onset (y)	43.54 (8.81)	46.64 (7.20)	41.65 (9.30)	0.106
Disease duration[a] (y)	7.38 (3.89)	6.86 (4.19)	7.70 (3.76)	0.387

[a]Measured at treatment initiation.
[b]Mann-Whitney test for quantitative data and chi square or Fisher exact test for qualitative data; y: in years; CAG: Cytosine-Adenine-Guanine; Responders and non-responders were defined by the CLEB algorithm.
Note: Quantitative data are expressed in mean (SD) and qualitative data in N(%).

Patients with high CAG repeats are more frequently non-responders than those with low CAG repeats (Table 1). The repetition of CAG is correlated with the disease being more serious.[28–30] The clustering results suggest that APRs are inefficient for patients with high CAG repeats (the patients with the most severe symptoms). The two profiles of evolution we observed could reflect the

Table 2. Concordance of responders and non-responders according to CLEB and LCMM methods.

	Responders LCMM	Non-responders LCMM
Responders CLEB	5	10
Non-responders CLEB	2	22

Note: CLEB: Clustering in Longitudinal data with Extended Baseline; LCMM: Latent-Class Mixed Model.

treatment effect, the disease severity or both. The conclusion could only be speculative and a confirmatory analysis is required.

Repeating the method on the 741 subsets of the whole data set that were created by deleting two patients each time showed that the identified subgroups are robust, with only four patients being classified less than 90% of the time in the same subgroup.

The LCMM method was also applied to the data set. Sixty nine percentage of patients had a matching classification with both the CLEB and the LCMM methods (Table 2). The conclusions were similar with non-responders having higher CAG repeats (Mann-Whitney test, $P = 0.020$).

5.2 The impact of bariatric surgery on BMI

Obesity is an abnormal accumulation of body fat. It is associated with increased health problems, such as hypertension, Type II diabetes, coronary disease and hyperlipidemia. The Body Mass Index (BMI), obtained by dividing the weight by the square of the height, quantifies the tissue mass in an individual. Obesity is defined as a BMI score higher than 30. Currently, there are three categories of treatment: dietary modification, medication and surgery. Surgery treats people with potentially life-threatening obesity when other treatments, such as lifestyle changes, have not worked. Here, we evaluated two types of bariatric surgeries: sleeve gastrectomy and gastric bypass. Data were obtained from the records of a French bariatric centre. In the current clustering study, we analysed the period of 12 months before treatment initiation and 12 months after to assess the effect of the treatment on weight loss before stabilization. Only those 39 women with at least one measurement before surgery, one measurement at surgery and one measurement after surgery were included. They were followed for an average of 15.50 months (SD = 4.51), representing between 3 and 8 visits.

We applied the CLEB method (with the model-based algorithm and V parameterization) to the 39 women suffering from obesity. The CLEB split the population into two subgroups: 18 high-responders and 21 low-responders to surgery. We then modelled the data with these groups of treatment responses as the covariate. The model showed a stabilization of the BMI during the pre-operative period (mean of slope: –0.06 points per month, $SE = 0.06$, $P = 0.350$) in the whole cohort. The difference in the slope between the baseline phase and the treatment effect phase was –1.34 points per month ($SE = 0.11$, $P < 0.001$) for high-responders and -0.75 points per month ($SE = 0.10$, $P < 0.001$) for low-responders. Low-responders had a lower BMI at treatment initiation, with a BMI of of 8.58 points ($SE = 1.90$, $P < 0.001$) less than high-responders.

Younger women with a high weight at surgery are more frequently classified in the group of high-responders (Table 3). This is consistent with the fact that pre-operative BMI is positively associated with weight loss over a short follow-up period after bariatric surgery, whereas the correlation becomes negative over a longer follow-up period.[31] Moreover, younger patients might lose more weight because of their high metabolic activity compared to older patients.[32]

Table 3. Description of high-responders and low-responders to surgery.

	Whole cohort $N = 39$	High-responders $N = 18$	Low-responders $N = 21$	*p*-values[b]
Age[a] (y)	44.91 (10.11)	38.86 (7.91)	50.09 (8.93)	<0.001
Treatment				
Sleeve	19 (48.72%)	8 (44.44%)	11 (52.38%)	0.621
Bypass	20 (51.28%)	10 (55.56%)	10 (47.62%)	
Weight[a]	114.10 (20.76)	127.56 (21.67)	102.57 (10.80)	<0.001
Type 2 diabetes				
Yes	6 (16.67%)	3 (16.67%)	3 (16.67%)	>0.999
No	30 (83.33%)	15 (83.33%)	15 (83.33%)	
NA	3	0	3	
Sleep apnea				
Yes	7 (18.92%)	4 (22.22%)	3 (15.79%)	0.693
No	30 (81.08%)	14 (77.78%)	16 (84.21%)	
NA	2	0	2	
Hypertension				
Yes	9 (25.00%)	2 (11.11%)	7 (38.89%)	0.121
No	27 (75.00%)	16 (88.89%)	11 (61.11%)	
NA	3	0	3	

Note: Quantitative data are expressed in mean (SD) and qualitative data in N(%).
[a]Measured at treatment initiation.
[b]Mann-Whitney test for quantitative data and chi square or Fisher exact test for qualitative data; y: in years; NA: Not Available; High-responders and low-responders were defined by the CLEB algorithm.

BMI at surgery is linked to treatment effect, so we performed a multivariate clustering as a sensitivity analysis. We applied the CLEB method (with the model-based algorithm and VVI parameterization) on $(\hat{b}_0, \hat{b}_2)$. The match between the univariate and the multivariate was 82% and the conclusions were similar, with a faster weight loss in younger women with higher weight and BMI at surgery initiation.

We also made a clustering that included 24 months post-surgery observations by changing the time into $log(\text{time} + 1)$ to avoid the non-linearity caused by the plateau in the stabilization period. Once again, the results match the previous analyses whether the clustering was univariate or multivariate.

6 Discussion

6.1 The CLEB algorithm with the EM model-based (V parameterization) strategy

In this paper, we have presented the CLEB method, a new method for classifying patients according to treatment efficacy in the case of continuous longitudinal data. The method has two steps. The first consists of modelling the entire trajectory of the data with measurements before and after treatment initiation. An extended baseline was used: data were modelled with two slopes, corresponding to the baseline and the treatment effect phases. The slope change appears at the assumed time of the

treatment effect. Thus, the model has three mixed components: intercept ($\beta_0 + b_{0i}$), slope during baseline phase ($\beta_1 + b_{1i}$), and difference between the slopes at the baseline phase and the treatment effect phase ($\beta_2 + b_{2i}$), where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is the vector of the fixed effects and $\boldsymbol{b_i} = (b_{0i}, b_{1i}, b_{2i})$ is the vector of the random effects for patient i. In the second step, the clustering is made on random predictions of $\boldsymbol{b_2}$ using the EM model-based algorithm assuming different variances (V parameterization) and allowing different shapes between clusters. The current study showed that the CLEB algorithm is useful for clustering patients according to treatment efficacy in the case of longitudinal data such as data obtained for a progressive disease (e.g. Huntington's disease[26]). The lag between treatment initiation and treatment effect has to be specified as an input parameter, but simulation studies showed that increase the variability of the lag does not have an impact on the results. This method is robust, regardless of the noise on the within- or between-subject variability of the baseline phase. Furthermore, the mixed model in the first step makes to the method insensitive to heterogeneity in the number and time of records between subjects. Even if the method could deal with missing data, patients need to have at least one measurement after treatment initiation to be included in the analysis. However, there is a minimum number of time points required to make the method efficient, and there also must be time points before the treatment initiation if the natural disease progression is heterogeneous.

6.2 The lack of relevance of the other strategies

We considered other clustering strategies in the second step of the CLEB method. The simulation study showed that multivariate clustering (on $(\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_2)$, $(\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$ or $(\hat{\boldsymbol{b}}_0, \hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2)$) could be preferred to univariate clustering (on $\hat{\boldsymbol{b}}_2$) only if the treatment effect was linked to the patient's baseline conditions and the variance was low, whether, parametric or non-parametric strategies are used. However, when the variability is high, multivariate clustering strategies add more noise, and univariate strategies must be preferred.

For the non-parametric strategies, partitioning algorithms were more relevant than AHC algorithms, and the Canberra distance provided better results than the Euclidean one with two balanced subgroups but was inefficient for unbalanced subgroups. Indeed, in the case of univariate clustering, this distance will always separate positive and negative values.

In unbalanced scenarios, all the methods, except for the CLEB method with a model-based algorithm strategy, failed when $N_A = 2$. Even though this case seemed unrealistic, it was considered because of the possibility that a treatment had a beneficial effect only for a rare genetic profile.

6.3 Comparison of CLEB and LCMM

Furthermore, the CLEB method was compared to the LCMM method. For all the simulation scenarios, the CLEB method performed as well as or better than the LCMM method, especially when there was high variability of the slope before treatment initiation. Indeed, with the LCMM method, the definition of the subgroups and the estimation of the parameters were done simultaneously, leading to a greater influence of the baseline slope on the subgroup definition. For a large variance in treatment efficacy, both the CLEB and the LCMM methods became inefficient. Indeed, for large variance, the distribution of random terms, which is a mixture of K Gaussian distributions, tended to become unimodal.[33]

6.4 Some extensions for the CLEB method

For all the simulation studies, the number of clusters was an input parameter for the CLEB method. The choice of two clusters may make the model find two distinct subgroups even if they do not exist. However, the Bayesian Information Criterion may help in the choice of the number of clusters.

The CLEB method could have some extensions with more specific models. Indeed, we considered the case of a sustainable treatment effect, using a piecewise linear mixed model with two slopes. However, it was easy to extend this to the case of a piecewise linear mixed model with three slopes, the third corresponding to a plateau in treatment efficacy or to a resumption of the disease progression. Thus, subgroups of patients were defined according to short- and long-term treatment efficacy.

Furthermore, the assumption of a linear constant change for the outcome could be false. Indeed, the CLEB method proposed the use of a linear mixed model which assumed a constant change for the outcome. This assumption may not hold for psychometric scores characterized by upper and lower bounds. Considering the outcome as a discrete and bounded variable can improve the model and classify patients better. Indeed, it has been shown that, for handling this type of data, an alternative mixed model, handling this type of data, performed better than the classical linear mixed models[34] in data modelling. Splines or wavelets are also some modelling alternatives for specific outcomes such as time series data.[35]

Finally, only unsupervised algorithms, which attributed each patient to a cluster without a prior subgroup, were envisaged. However, if some patients could be easily identified as treatment responders or non-responders, with a mixture of labelled and unlabelled data, the algorithm would be improved by training it on the labelled patients and then applying it to the unlabelled patients as a semi-supervised algorithm.

6.5 Perspectives

This new method will help to define subgroups in the search for markers of treatment efficacy and to understand why some patients respond to treatment, while others fail to do so. It extracts information from pharmaco-epidemiological studies (the treatment arm of clinical trials or cohort studies with a treatment initiation during the follow-up). It is particularly interesting to find small subgroups of responders to a treatment that has never demonstrated its efficacy in a clinical trial. The definition of subgroups may help to find marker(s) of treatment response, which is a prerequisite for the implementation of stratified design for future clinical trials. This leads to therapies being matched with a specific patient population. It is anticipated that this will have a major effect on both clinical practice and the development of new drugs and diagnostics.[36]

Acknowledgments

The authors thank Désaméricq G., Dolbeau G., Verny C., Charles P., Durr A., Youssov K., Simonin C., Azulay J-P., Tranchant C., Goizet C., Damier P., Broussolle E., Demonet J-F., Morgado G., Cleret de Langavant L., Macquin-Mavier I. and Maison P. for their data of patients with Huntington's disease receiving antipsychotics and related drugs and Lazzati A. for his data of patients suffering from obesity and having a bariatric surgery. The authors thank Bardy P. from Centre des langues of Paris Descartes University and the proofreaders from Proof-Reading-Service.com for their english language correction.

The authors thank the neurologists and the neuropsychologists from the Multicentric Intracerebral Grafting in Huntington's Disease trial who collected the data for the simulation study: A-C. Bachoud-Lévi, MD, PhD

(Henri Mondor hospital, Créteil, Principal investigator), M-F Boissé (Henri Mondor hospital, Créteil, Neuropsychologist), L. Lemoine (Henri Mondor hospital, Créteil, Neuropsychologist), C. Verny, MD, PhD (Angers hospital, Site coordinator), G. Aubin (Angers hospital, Neuropsychologist), J-F Demonet, MD, PhD (CHU Rangueil, Toulouse, Site coordinator), F. Calvas, MD (CHU Rangueil, Toulouse, Investigator), P. Krystkowiak, MD, PhD (Roger Salengro hospital, Lille and and CHU d'Amiens, Amiens, Sites coordinator), C. Simonin, MD, PhD (Roger Salengro hospital, Lille, Investigator), M. Delliaux (Roger Salengro hospital, Lille, Neuropsychologist), P. Damier, MD, PhD (Nord Laennec hospital, Nantes, Site coordinator), P. Renou (Nord Laennec hospital, Nantes, Investigator), F. Supiot (Erasme hospital, Bruxelles, Site coordinator), H. Slama (Erasme hospital, Bruxelles, Neuropsychologist).

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Mrs. Schramm was successively supported by the NeuroStemcell Consortium (European Community Seventh Framework Program grant Agreement No. 222943) and by ''Investments for the future'' (ANR-11-INBS-0011 - NeurATRIS: Infrastructure de recherche translationnelle pour les biothrapies en Neurosciences).

Dr. Katsahian reports no disclosures.

Dr. Vial reports no disclosures.

Dr. Bachoud-Lévi was consultant for Teva once in 2014. She received grants from the Ministry of Health supporting the National reference center for Huntington's disease and several grants for academic trials provided by the ''Direction de la Recherche Clinique'' (AP-HP). She is partner from several investments for the future (Labex IEC, Neuratris) and a FP7 EU grant FP7 (Repair-HD).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by investment for the future NeurATRIS: Infrastructure de recherche translationnelle pour les biothrapies en Neurosciences (ANR-11-INBS-0011), European Community Seventh Framework Program Neurostemcell (Grant Agreement no. 222943), European Community Seventh Framework Program Repair-HD (Grant Agreement no 602245). The Département d'Études Cognitives of the École Normale Supérieure is supported by two ANR grants from the French Research Agency (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL).

References

1. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*, 5th ed. Chichester, West Sussex, U.K: Wiley-Blackwell, 2011.
2. Koestler DC, Marsit CJ, Christensen BC, et al. A recursively partitioned mixture model for clustering time-course gene expression data. *Trans Cancer Res* 2014; **3**: 217.
3. Harrington M, Velicer WF and Ramsey S. Typology of alcohol users based on longitudinal patterns of drinking. *Addict Behav* 2014; **39**: 607–621.
4. Castellini G, Fioravanti G, Sauro CL, et al. Latent profile and latent transition analyses of eating disorder phenotypes in a clinical sample: a 6-year follow-up study. *Psych Res* 2013; **207**: 92–99.
5. Kent P and Kongsted A. Identifying clinical course patterns in sms data using cluster analysis. *Chiropract Manual Therap* 2012; **20**: 1–12.
6. Tepper PG, Randolph JF Jr, McConnell DS, et al. Trajectory clustering of estradiol and follicle-stimulating hormone during the menopausal transition among women in the study of women's health across the nation (swan). *J Clin Endocrinol Metabol* 2012; **97**: 2872–2880.
7. Muthén B and Muthén LK. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clin Exp Res* 2000; **24**: 882–891.
8. Verbeke G and Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *J Am Stat Assoc* 1996; **91**: 217–221.
9. Genolini C and Falissard B. Kml: k-means for longitudinal data. *Comput Stat* 2010; **25**: 317–328.
10. Madsen K, Miller J and Province M. The use of an extended baseline period in the evaluation of treatment in a longitudinal duchenne muscular dystrophy trial. *Stat Med* 1986; **5**: 231–241.

11. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963–974.
12. Laird NM. Missing data in longitudinal studies. *Stat Med* 1988; **7**: 305–315.
13. Fraley C and Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; **97**: 611–631.
14. Redner RA and Walker HF. Mixture densities, maximum likelihood and the em algorithm. *SIAM Rev* 1984; **26**: 195–239.
15. Banfield JD and Raftery AE. Model-based gaussian and non-gaussian clustering. *Biometrics* 1993; **49**: 803–821.
16. Hartigan JA and Wong MA. Algorithm as 136: a k-means clustering algorithm. *Appl Stat* 1979; **28**: 100–108.
17. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, pp. 281–297.
18. Kaufman L and Rousseeuw P. Statistical data analysis based on the Ll norm. In: dodge Y (ed.) *Clustering by means of medoids*. Amsterdam: North-Holland, 1987, pp.405–416.
19. Day WH and Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *J Class* 1984; **1**: 7–24.
20. Team RDC. R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria, http://www.R-project.org (2013).
21. Pinheiro J, Bates D, DebRoy S, et al. nlme: linear and nonlinear mixed effects models, http://CRAN.R-project.org/package = nlme. R package version 3.1-105 (2012).
22. Fraley C, Raftery AE, Murphy TB, et al. mclust: normal mixture modeling for model-based clustering, classification, and density estimation, http://CRAN.R-project.org/package=mclust. R package version 4.3 (2012).
23. Lucas A. *amap: another multidimensional analysis package*, http://CRAN.R-project.org/package=amap. R package version 0.8-12 (2014).
24. Maechler M and Rousseeuw P. *cluster: cluster analysis basics and extensions*, http://CRAN.R-project.org/package=cluster. R package version 1.14.3 (2012).
25. Proust-Lima C, Philipps V, Diakite A, et al. *lcmm: estimation of extended mixed models using latent classes and latent processes*, http://CRAN.R-project.org/package=lcmm. R package version 1.6.6 (2014).
26. Bates G, Harper P and Jones L. *Huntington's disease: oxford monographs on medical genetics*. New York: Oxford University Press, 2002.
27. Désaméricq G, Dolbeau G, Verny C, et al. Effectiveness of anti-psychotics and related drugs in the huntington french-speaking group cohort. *PLoS ONE* 2014; **9**: e85430. DOI:10.1371/journal.pone.0085430.
28. Trottier Y, Biancalana V and Mandel JL. Instability of cag repeats in huntington's disease: relation to parental transmission and age of onset. *J Med Gen* 1994; **31**: 377–382.
29. Brandt J, Bylsma F, Gross R, et al. Trinucleotide repeat length and clinical progression in huntington's disease. *Neurology* 1996; **46**: 527–531.
30. Langbehn DR, Hayden MR and Paulsen JS. Cag-repeat length and the age of onset in huntington disease (hd): a review and validation study of statistical approaches. *Am J Med Gen Part B: Neuropsych Gen* 2010; **153**: 397–408.
31. Livhits M, Mercado C, Yermilov I, et al. Preoperative predictors of weight loss following bariatric surgery: systematic review. *Obes Surg* 2012; **22**: 70–89.
32. Agüera Z, García-Ruiz-de Gordejuela A, Vilarrasa N, et al. Psychological and personality predictors of weight loss and comorbid metabolic changes after bariatric surgery. *European Eating Disorders Rev* 2015; **23**: 509–516.
33. Strenio JF, Weisberg HI and Bryk AS. Empirical bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* 1983; **39**: 71–86.
34. Proust-Lima C, Dartigues JF and Jacqmin-Gadda H. Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *Am J Epidemiol* 2011; **174**: 1077–1088.
35. James GM and Sugar CA. Clustering for sparsely sampled functional data. *J Am Stat Assoc* 2003; **98**: 397–408.
36. Trusheim MR, Burgess B, Hu SX, et al. Quantifying factors for the success of stratified medicine. *Nature Rev Drug Discov* 2011; **10**: 817–833.

3.2 Simulations supplémentaires

3.2.1 Estimation de la différence d'effet entre les deux groupes

Afin d'évaluer les estimations de différence entre les deux sous-groupes, nous avons appliqué le modèle suivant sur les données simulées pour l'article :

$$y_{ij} = (c_0 + c_1 \times g_i) + (c_2 + c_3 \times g_i) \times t_{ij} + (c_4 + c_5 \times g_i) \times (t_{ij} - \tau) \times \mathbb{1}(t_{ij} \geq \tau) + \gamma_i + \varepsilon_{ij} \quad (3.2)$$

où y_{ij} est le score du patient i lors de la $j$ème visite au temps t_{ij}, τ est le délai entre l'initiation du traitement et sa prise d'effet, γ_i est un terme aléatoire au niveau du patient et $g_i \in \{A,B\}$ est le sous-groupe auquel appartient le patient i, estimé avec notre méthode CLEB. $c = (c_0...c_5)$ est le vecteur des coefficients de régression du modèle, avec c_5, le coefficient d'intérêt représentant la différence de l'effet de traitement entre les deux sous-groupes A et B.

Nous avons estimé c_5 pour différents scénarios et présentons les résultats obtenus dans le cas où nous faisons varier $\mu_2^{(A)}$ et $\mu_2^{(B)}$. Les résultats sont présentés en moyenne et écart-type des estimations obtenues sur les 1000 simulations. Nous donnons, de plus, le pourcentage de fois où la p-valeur associée au test de Wald pour le coefficient c_5 ($H_0 : c_5 \neq 0$) est inférieure à 0,05. Les résultats sont présentés dans la table 3.

Table 3 – Estimation de c_5 et puissance associée au test de Wald dans le cas de deux sous-groupes pour différentes valeurs de $\mu_2^{(A)} - \mu_2^{(B)}$

	$\mu_2^{(A)} - \mu_2^{(B)}$				
	-2	-4	-6	-8	-10
CLEB					
Modèle de mélange paramétrisation V	-2,77 (1,10) 90,50%	-4,28 (0,74) 99,50%	-6,05 (0,64) 100%	-8,02 (0,63) 100%	-10,01 (0,63) 100%
Modèle de mélange paramétrisation E	-3,16 (0,97) 98,46%	-4,33 (0,69) 100%	-6,07 (0,63) 100%	-8,02 (0,63) 100%	-10,01 (0,63) 100%
K-moyenne distance euclidienne	-2,88 (0,75) 98,67%	-4,27 (0,64) 100%	-6,07 (0,63) 100%	-8,02 (0,63) 100%	-10,01 (0,63) 100%
K-moyenne distance de Canberra	-2,81 (0,71) 99,18%	-4,25 (0,64) 100%	-6,06 (0,62) 100%	-8,02 (0,63) 100%	-10,01 (0,63) 100%
LCMM	-1,84 (0,63) 91,48%	-3,46 (0,67) 100%	-5,55 (0,67) 100%	-7,76 (0,67) 100%	-9,92 (0,65) 100%

Ces résultats ont été obtenus pour $N_A = N_B = 25$.

3.2.2 Estimation du nombre de clusters par le critère d'information bayésien (BIC)

Nous avons simulé des données avec un seul groupe de patients ou avec deux sous-groupes en faisant varier la différence entre les groupes. Nous avons réalisé le clustering avec la méthode CLEB associée à l'algorithme paramétrique supposant un modèle de mélange fini gaussien avec une paramétrisation V (variance inégale entre les sous-groupes). Nous avons utilisé le critère BIC pour estimer le nombre de sous-groupes optimal. Les résultats sont exprimés en pourcentage de fois où le nombre de clusters est choisi sur les 1000 simulations et sont résumés dans la table 4.

TABLE 4 – Nombre de clusters défini par le critère BIC

	Nombre de clusters défini par le critère du BIC				
	1	2	3	4	≥ 5
Données simulées					
1 cluster de 50 patients	97,12%	2,45%	0,32%	0,11%	0%
2 clusters de 25 patients chacun avec					
$\lvert\mu_2^{(A)} - \mu_2^{(B)}\rvert = 10$	0%	97,38%	2,12%	0,50%	0%
$\lvert\mu_2^{(A)} - \mu_2^{(B)}\rvert = 8$	0%	97,18%	2,62%	0,20%	0%
$\lvert\mu_2^{(A)} - \mu_2^{(B)}\rvert = 6$	29,38%	67,30%	3,12%	0,20%	0%
$\lvert\mu_2^{(A)} - \mu_2^{(B)}\rvert = 4$	88,91%	10,38%	0,71%	0%	0%

Lorsqu'il n'y a qu'un seul groupe de patients (traitement avec efficacité identique pour tous), le critère BIC identifiera artificiellement au moins deux sous-groupes distincts dans 2,88% des cas uniquement. Lorsqu'il existe deux sous-groupes distincts de réponse au traitement, mais que l'écart entre les sous-groupes est faible, le critère BIC a tendance à sous-estimer le nombre de clusters.

3.2.3 Comparaison avec la méthode KML

Nous avons comparé notre méthode avec la méthode KML pour des scénarios où les délais inter-mesures sont identiques. Nous avons appliqué la méthode KML grâce au package R `kml` [93] et l'avons paramétrée avec les distances euclidienne et de Canberra.

Tout comme les méthodes CLEB et LCMM, la méthode KML répartit correctement les patients au sein des deux sous-groupes quelles que soit la variabilité du délai entre l'initiation du traitement et le début de son effet et le nombre de patients. Tout comme les

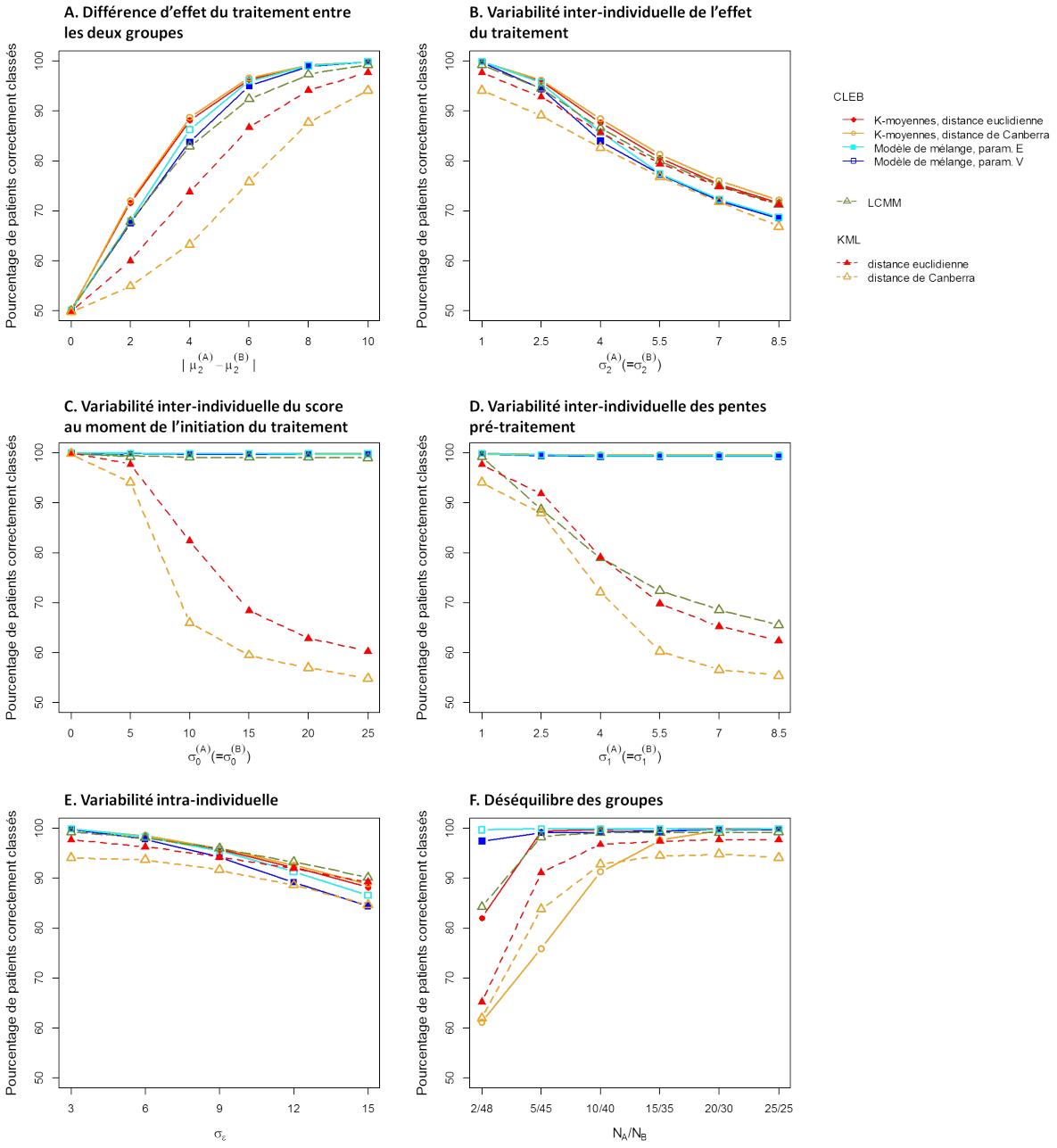


FIGURE 17 – Pourcentage de patients correctement classés avec la méthode KML

méthodes CLEB et LCMM, la méthode KML est moins performante lorsque la variabilité inter-individuelle de l'effet du traitement ou la variabilité intra-individuelle augmentent (Figure 17.B et E) ou la différence d'effet du traitement entre les deux groupes diminue (Figure 17.A). Lorsque la variabilité inter-individuelle du score au moment de l'initiation du traitement augmente, la performance de la méthode KML diminue (Figure 17.C). En effet, dans la méthode KML, le score moyen a un poids plus important que la forme de la trajectoire dans le calcul de la distance entre deux patients. Une augmentation de la

variabilité des scores au moment de l'initiation des patients entraîne une translation des trajectoires selon l'axe des ordonnées et donc un possible éloignement des patients du même groupe selon cet axe. De même lorsque la variabilité inter-individuelle des pentes pré-traitement augmente, les résultats de KLM sont similaires à ceux obtenus avec la méthode LCMM (Figure 17.D). Là encore l'augmentation de cette variabilité entraîne une translation selon l'axe des ordonnées des premiers et derniers points de la trajectoire. Enfin, lorsque les groupes sont déséquilibrés, la méthode n'est impactée qu'en cas d'un sous-groupe à très faible effectif (Figure 17.F).

3.2.4 Comparaison avec la méthode par régressions individuelles

Nous avons comparé notre méthode avec un clustering réalisé sur les coefficients issus de régressions individuelles. Cette méthode se décompose aussi en deux étapes. Premièrement, pour chaque patient, nous avons estimé les coefficients d'un modèle de régression linéaire par morceaux selon l'équation (3.3).

$$y_j^{(i)} = \beta_0^{(i)} + \beta_1^{(i)} \times t_j^{(i)} + \beta_2^{(i)} \times (t_j^{(i)} - \tau) \times \mathbb{1}(t_{j^{(i)}} \geq \tau) + \varepsilon_j^{(i)} \tag{3.3}$$

où $\beta^{(i)} = (\beta_0^{(i)} \beta_1^{(i)} \beta_2^{(i)})$ est le vecteur des coefficients de régression associés à l'individu i. τ est le délai entre l'initiation du traitement et son effet. Deuxièmement, le clustering a été réalisé sur le coefficient de régression correspondant au changement de pente ($\beta_2^{(i)}$), grâce aux méthodes de clustering pour données transversales (algorithme des K-moyennes avec distance euclidienne ou de Canberra et algorithme basé sur un modèle de mélange fini avec hypothèse de variances égales ou inégales). Nous avons comparé les résultats obtenus par cette méthode et ceux obtenus avec la méthode CLEB (Figure 18).

Nos résultats montrent que les deux méthodes donnent des résultats similaires excepté lorsque la variabilité intra-patient est élevée (Figure 18.C). Plus la variabilité intra-patient augmente, plus la méthode CLEB aura de meilleures performances que la méthode par régressions individuelles. En effet, dans la méthode par régressions individuelles, les coefficients associés à chaque patient sont estimés à partir d'un faible nombre de données et donc leur estimation est biaisée en cas de forte variabilité des données. Au contraire, l'algorithme CLEB utilise un modèle sur l'ensemble des patients ce qui permet de « lisser » ces variations et donc d'y être moins sensible.

Nous pouvons noter que lorsque les groupes sont déséquilibrés, l'algorithme des K-moyennes avec la distance de Canberra donne de meilleurs résultats au sein de la méthode par régressions individuelles qu'au sein de la méthode CLEB (Figure 18.D). En effet, avec une seule variable en entrée, la distance de Canberra sépare les valeurs positives des valeurs négatives. Cela avantage cette méthode dans la mesure où nous avons simulé des données

avec des changements de pente positifs pour le groupe « non répondeurs » et négatifs pour le groupe « répondeurs ».

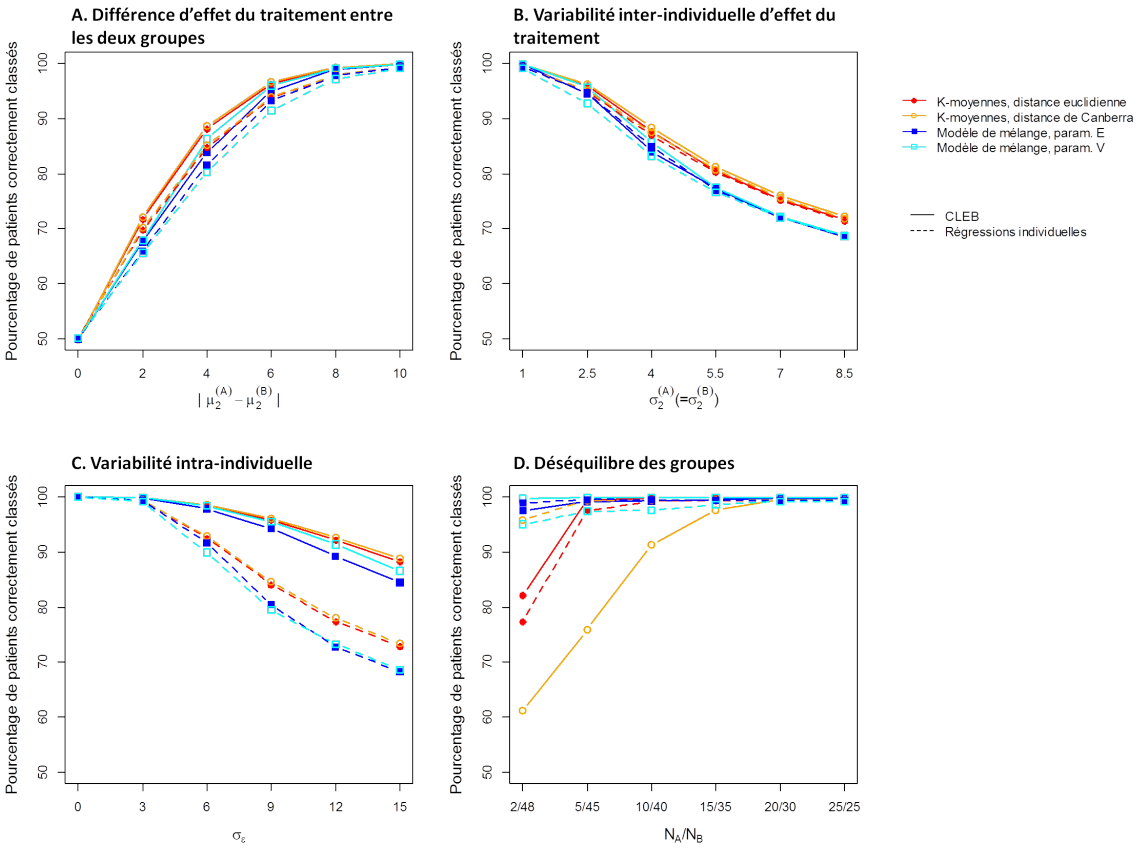


FIGURE 18 – Pourcentage de patients correctement classés avec la méthode par régressions individuelles

3.3 Application à MIG-HD

Dans l'étude MIG-HD, des sous-groupes de patients « répondeurs » et « non répondeurs » à la greffe ont été définis en comparant l'évolution post-greffe des performances motrices (test moteur de l'UHDRS) de chaque patient à l'évolution standard d'un groupe contrôle. Le groupe contrôle est constitué de 45 patients, issus de la cohorte RHLF, appariés pour la visite correspondant à la date de la greffe aux 45 patients greffés de MIG-HD. L'appariement a été réalisé sur l'âge, le sexe, le nombre de répétitions de CAG, la durée de la maladie, le score moteur à la visite d'appariement et la pente d'évolution du score moteur avant la visite d'appariement (qui correspond à la pente pré-greffe des patients MIG-HD). Soit β_{RHLF} la pente d'évolution moyenne du groupe contrôle et $95\%CI_{\text{RHLF}}$ son intervalle de confiance. Les patients greffés de MIG-HD ayant une pente post-traitement inférieure à $min(95\%CI_{\text{RHLF}})$ sont considérés comme « répondeurs », ceux ayant une

pente post-traitement supérieure à $max(95\%CI_{\text{RHLF}})$ sont considérés comme « non répondeurs ». Les patients intermédiaires ont été analysés comme des « non répondeurs ».

Nous proposons ici d'appliquer notre méthode CLEB aux données de MIG-HD dans le but de définir deux sous-groupes de patients. Les effets aléatoires présentés sur la figure 19 sont issus du modèle linéaire mixte à deux pentes appliqué sur les données de MIG-HD.

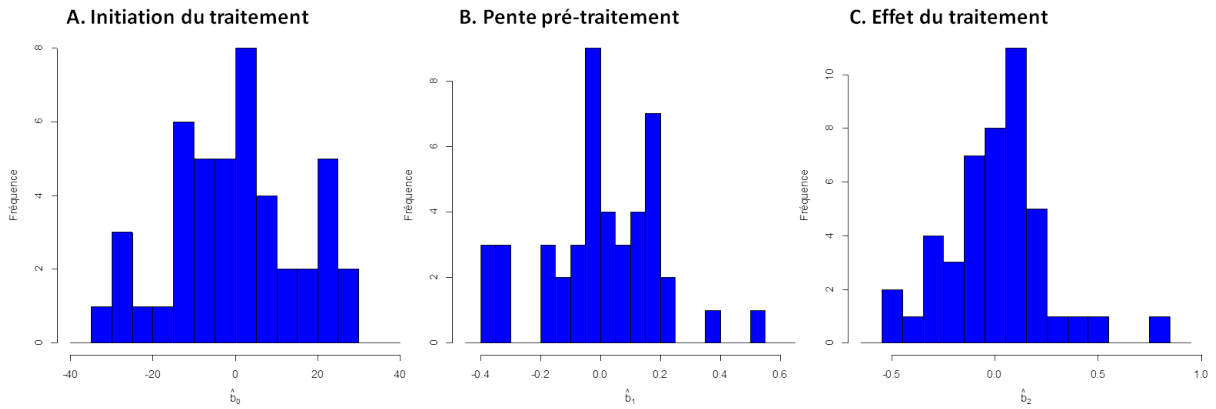


FIGURE 19 – Distribution des effets aléatoires du modèle mixte à deux pentes appliqué sur les données de MIG-HD

Analyse univariée sur $\hat{b}_2$:

En appliquant l'algorithme de clustering paramétrique (paramétrisation V : variance différente entre les clusters), le critère du BIC indique que le nombre optimal de clusters se réduit à un seul. En effet la distribution des effets aléatoires $\hat{b}_2$ semble unimodale (Figure 19.C).

Cependant en présence de fortes variabilités, les mélanges de lois sont difficiles à déterminer car ils se rapprochent d'une loi normale unimodale [94], ce qui peut être notre cas dans MIG-HD. Nos simulations montrent en effet que l'augmentation de la variabilité de l'effet du traitement diminue le pourcentage de patients bien classés par l'algorithme CLEB, quelle que soit la stratégie envisagée à la seconde étape. De plus, nous avons montré que dans le cas de faible différence entre les groupes, le BIC ne permet plus de déterminer le nombre de sous-groupes optimal.

En forçant le modèle à trouver deux clusters de patients, nous trouvons deux profils semblables (Figure 20). L'analyse univariée n'a pas mis en évidence plusieurs profils de réponse à la greffe.

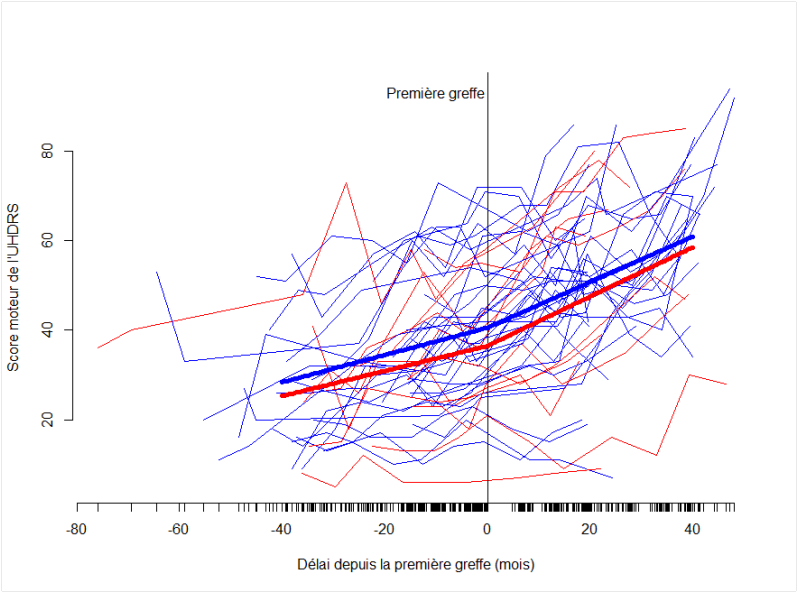


FIGURE 20 – Définition des sous-groupes par l'analyse univariée sur $\hat{b}_2$

Les patients du groupe A ont un nombre de voxels hypométaboliques dans le striatum plus faible que les patients du groupe B, ce qui signifie que les patients du groupe A ont une atrophie plus grande. Le score « performances motrices et fonctionnelles » pré-traitement correspond à la dimension motrice/fonctionnelle d'une analyse en composantes principales. Un score plus élevée correspond à une vitesse de déclin sur le plan moteur/fonctionnel plus rapide.

Analyse multivariée sur $\hat{b}_0$,$\hat{b}_1$,$\hat{b}_2$:

Les distributions de $\hat{b}_0$ et de $\hat{b}_1$ ne semblant pas unimodales (Figure 19.A et C), nous avons aussi réalisé une analyse multivariée. En appliquant l'algorithme de clustering paramétrique (paramétrisation VVI : variances différentes entre les clusters et entre les effets aléatoires), le critère du BIC indique que le nombre optimal de clusters est de deux.

Cette analyse met en avant non pas deux profils de réponse à la greffe mais deux stades de la maladie (Figure 21).

En effet, le groupe A regroupe les patients pour lesquels la maladie évolue plus vite ou qui sont à un stade plus avancé de la maladie que les patients du groupe B (Table 5).

TABLE 5 – Nombre de clusters défini par le BIC

	groupe A	groupe B	p-valeur
Voxels hypométaboliques (striatum)	1350,33 (503,39)	957,93 (419,36)	0,009
Nombre de répétitions de CAG	46,5 (4,89)	43,86 (1,92)	0,067
Score de capacité fonctionnelle (TFC)	9,73 (2,05)	10,80 (1,86)	0,068
Score « performances motrices et fonctionnelles » pré-traitement	0,49 (1,10)	−0,83 (0,81)	$< 0{,}001$

Les p-valeurs ont été obtenus par le test de Mann-Whitney et sont non corrigées.

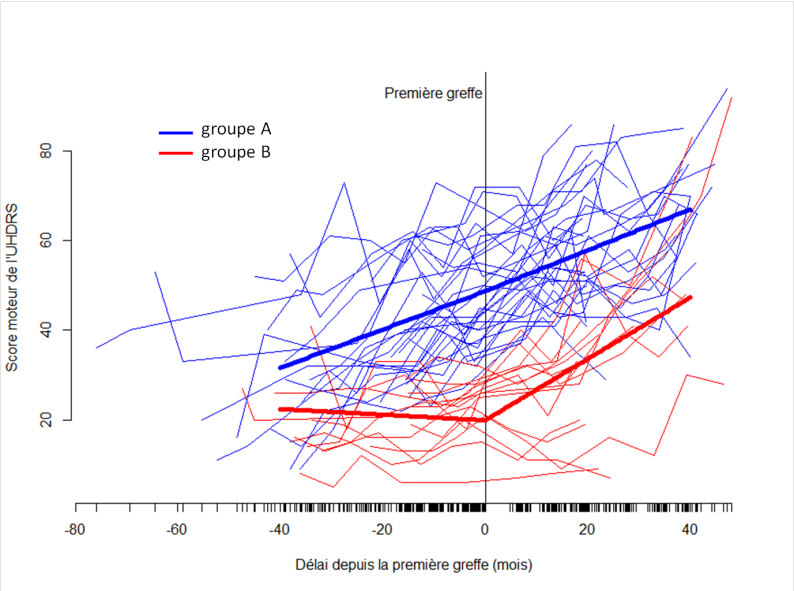


FIGURE 21 – Définition des sous-groupes par l'analyse multivariée sur $\hat{b}_0$, $\hat{b}_1$ et $\hat{b}_2$

Comparaison avec la méthode CLEB combinée à un algorithme non paramétrique

En appliquant l'algorithme non paramétrique des K-moyennes avec la distance euclidienne sur les effets aléatoires $\hat{b}_2$, nous trouvons deux sous-groupes différents en terme d'évolution pré- et post-traitement. Ces sous-groupes sont proches de ceux définis par la comparaison avec la cohorte RHLF (Table 6) avec 85% des patients non intermédiaires classés dans le même groupe.

TABLE 6 – Concordance entre les groupes définis par la comparaison avec les patients de la cohorte RHLF et ceux définis avec le CLEB (algorithme des K-moyennes et distance euclidienne)

	CLEB « répondeurs »	CLEB « non répondeurs »
RHLF « répondeurs »	14	6
RHLF « intermédiaires »	5	4
RHLF « non répondeurs »	0	16

Ces résultats montrent qu'il est possible de construire artificiellement des sous-groupes de « répondeurs » et de « non répondeurs » à la greffe et que ces sous-groupes sont proches de ceux construits par comparaison avec les patients issus de la cohorte RHLF. Cela montre que notre méthode permet de définir des sous-groupes de patients sans utiliser des données autres que celles des patients inclus dans l'analyse.

Discussion

La méthode CLEB avec la stratégie paramétrique basée sur les modèles de mélange, appliquée sur les données de MIG-HD, n'a pas pu mettre en évidence des différences de réponse à la greffe. Cependant, nous montrons, grâce à une analyse de clustering multivarié, qu'il existe deux profils de patients dans cet échantillon, différents en terme d'avancement de la maladie au moment de la greffe. Dans cette analyse, nous avons supposé que le changement de pente se réalisait au moment de l'initiation du traitement. Or, les études précédentes semblent indiquer un effet de la greffe à partir de 18 à 20 mois post-opération. Avec notre méthode nous aurions pu tenir compte de ce délai mais cela nécessite d'avoir plus de données après 20 mois de suivi post-greffe. Ces données pourront être disponibles grâce à l'étude « post MIG-HD ».

Deuxième partie

Intégration de nouveaux paramètres dans la conception des essais cliniques

Chapitre 4

Marqueurs pronostiques et marqueurs prédictifs (Etat de l'art)

4.1 Définition générale

Le *Biomarkers Definitions Working Group* a défini un marqueur en 2001 comme étant « une caractéristique qui est évaluée et mesurée objectivement comme un indicateur normal d'un processus biologique, pathologique, ou une réponse pharmacologique à une intervention thérapeutique » [95]. Pour être efficace, un marqueur doit être stable avec une technique de mesure fiable, reproductible, facile et rapide à mettre en œuvre, de préférence non invasive, avec un ratio temps-coût/bénéfice intéressant de façon à pouvoir être utilisé chez un grand nombre de patients. Deux grands types de marqueurs peuvent être identifiés : les marqueurs pronostiques et les marqueurs prédictifs. Un marqueur pronostique prédit le niveau de la maladie en l'absence de traitement. Un marqueur prédictif prédit l'effet du traitement. Un marqueur peut être à la fois pronostique et prédictif (Figures 22 et 23). Dans ce chapitre et les suivants, nous noterons M un marqueur pronostique et/ou prédictif binaire tel que $M+$ correspond à un marqueur positif et $M-$ à un marqueur négatif.

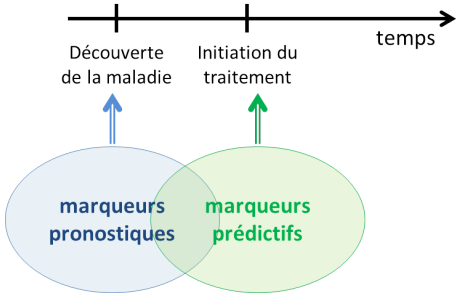


FIGURE 22 – Représentation schématique de l'impact des marqueurs prédictif et/ou pronostique

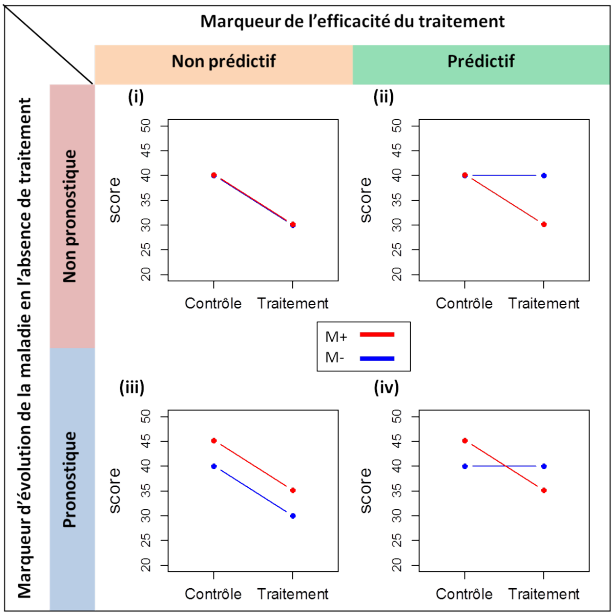


FIGURE 23 – Représentation schématique de l'impact des valeurs prédictive et pronostique d'un marqueur sur l'évolution de la maladie

Figure inspirée de [96]. Soit le marqueur M qui peut (i) n'avoir aucun impact, (ii) est prédictif seulement, (iii) est pronostique seulement ou (iv) est à la fois pronostique et prédictif. Dans notre exemple, un score élevé correspond à une maladie plus grave. Si le marqueur est pronostique, $M+$ a un score de base plus élevé que $M-$ et s'il est prédictif, $M+$ a un effet bénéfique du traitement tandis que $M-$ n'a pas d'effet du traitement. Lorsque M n'est pas pronostique, les deux groupes ont le même score de base. Lorsque M n'est pas prédictif, les deux groupes ont un effet bénéfique du traitement identique.

4.2 Définition dans le cadre d'une maladie évolutive

Dans une maladie évolutive telle que la maladie de Huntington, les marqueurs pronostiques vont se focaliser sur l'évolution de la maladie en l'absence de traitement (Figure 24) et les marqueurs prédictifs sur l'impact du traitement sur la pente du score modélisant l'évolution de la maladie (Figure 25). Connaître les paramètres biologiques, génétiques ou environnementaux qui régissent l'évolution d'une maladie peut aider à améliorer la prise en charge des patients et mieux comprendre les facteurs régulant la maladie. Un marqueur prédictif cliniquement efficace peut avoir un impact sur les essais cliniques, notamment en diminuant les effets secondaires. En ne donnant pas le traitement aux patients ne pouvant pas avoir un bénéfice de celui-ci, on améliore la balance bénéfice-risque.
Plus de 80% des maladies rares, comme la maladie de Huntington, sont causées par une anomalie génétique. Comprendre l'impact des mécanismes sous-jacents sur l'évolution de la maladie et la réponse au traitement est essentiel, non seulement pour ces maladies rares, mais aussi pour toutes les autres maladies plus courantes [97, 98].

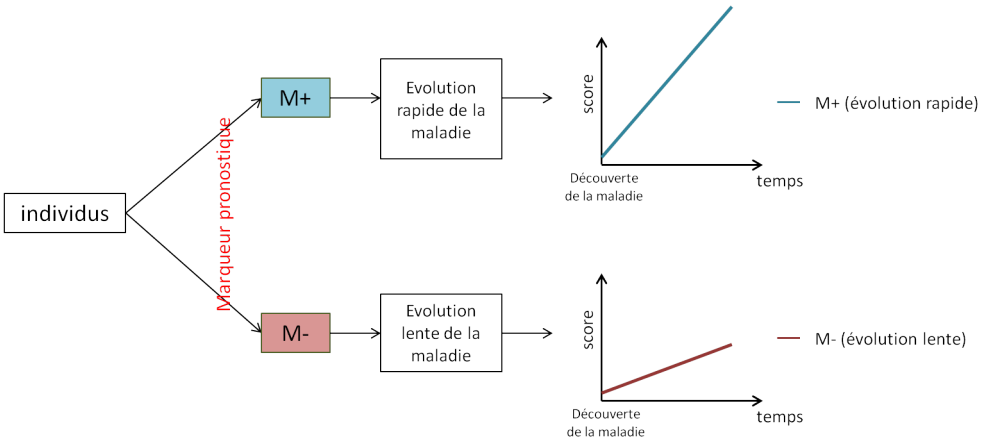


FIGURE 24 – Représentation schématique de l'évolution de la maladie en fonction du marqueur pronostique

Le score est une variable quantitative modélisant la progression de la maladie. Il augmente lorsque la maladie progresse. M représente un marqueur pronostique de l'évolution de ce score en l'absence de traitement. Si le patient appartient au sous-groupe $M+$, on prédit qu'il aura une progression rapide de la maladie et donc une augmentation rapide du score. Si le patient appartient au sous-groupe $M-$, on prédit qu'il aura une une progression lente de la maladie et donc une augmentation lente du score.

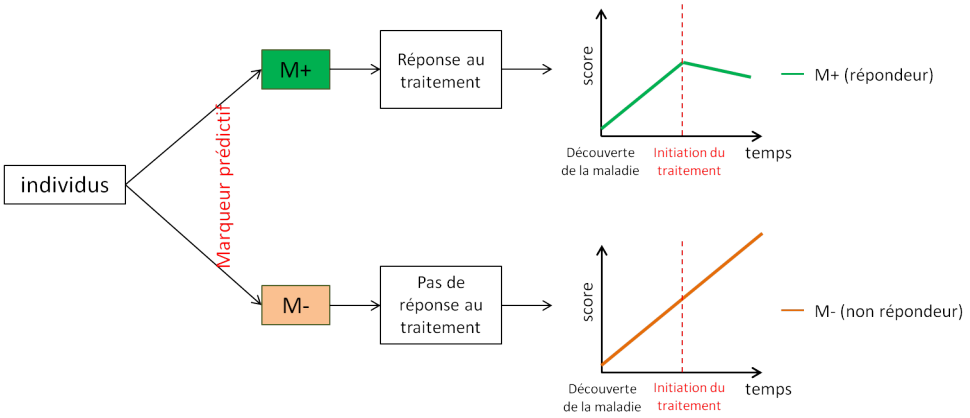


FIGURE 25 – Représentation schématique de l'impact du traitement sur l'évolution de la maladie en fonction du marqueur prédictif

Le score est une variable quantitative modélisant la progression de la maladie. Il augmente lorsque la maladie progresse. Au cours de leur maladie, les patients reçoivent un traitement. M représente un marqueur prédictif de l'efficacité de ce traitement. Si le patient appartient au sous-groupe $M+$, on prédit qu'il sera répondeur au traitement. Ainsi son score augmente jusqu'à l'initiation du traitement puis la pente d'évolution du score change (elle diminue, se stabilise ou continue à augmenter avec une plus faible pente). Si le patient appartient au sous-groupe $M-$, on prédit qu'il ne sera pas répondeur au traitement. Ainsi son score augmente jusqu'à l'initiation du traitement puis continue d'augmenter sans qu'il n'y ait d'impact du traitement sur l'évolution du score.

4.3 Utilisation des marqueurs en soins courants

En médecine, les marqueurs pronostiques et prédictifs peuvent être utilisés en soins courants. En effet, ils permettent de définir quel est le traitement le plus approprié à chaque patient. Le marqueur pronostique permettra d'identifier les patients avec une évolution lente ou rapide de la maladie, de mesurer la probabilité de rechute ou encore la gravité de la maladie. Cela implique de pouvoir choisir le traitement le mieux adapté en fonction du patient. Par exemple, si la probabilité de rechute est faible, le patient pourra bénéficier d'un traitement moins contraignant et/ou avec moins de toxicité qu'un patient ayant une forte probabilité de rechute sans traitement.

Les marqueurs prédictifs sont aussi directement utilisés pour choisir le traitement le mieux adapté au patient. Par exemple, ils permettent de ne pas donner un traitement à toxicité élevée à un patient qui ne pourrait pas en avoir un bénéfice.

Chapitre 5

Intégration des marqueurs pronostiques dans les essais cliniques : le polymorphisme COMT comme exemple de marqueur pronostique dans la maladie de Huntington

Les marqueurs pronostiques modulent l'évolution de la maladie en l'absence de traitement (Section 4.2). Par exemple, pour la maladie de Huntington, le nombre de CAG sur le gène IT15 du chromosome 4, en plus d'être un marqueur diagnostique (il détermine si l'individu est atteint de la maladie de Huntington ainsi que l'âge de début de la maladie [7]), est un marqueur pronostique. En effet, un nombre de répétitions de CAG plus élevé correspond à une progression de la maladie plus rapide [99, 100]. Cependant, d'autres facteurs génétiques inconnus peuvent moduler la maladie [101]. Nous avons étudié une mutation sur le gène cathecol-O-methyltransferase (COMT). Nous discutons ici de son impact sur la progression de la maladie de Huntington et comment nous pourrons en tenir compte dans les futurs essais cliniques.

5.1 Article « COMT Val^{158}Met Polymorphisms Modulates Huntington's Disease Progression »

Contexte

Le gène COMT dans le bras long du chromosome 22 (22q11) est connu pour réguler la dopamine dans le cerveau, en particulier dans le cortex préfrontal. L'un des polymorphismes les plus étudiés de ce gène est la mutation Valine (Val) en Méthionine (Met) à la position 158, soit le polymorphisme Val^{158}Met [102]. Chaque individu peut alors être soit homozygote Val/Val (Valine sur chaque allèle), soit homozygote Met/Met (Méthionine sur chaque allèle), soit hétérozygote Met/Val (Méthionine sur un allèle et Valine sur l'autre). L'activité du gène COMT est 38% plus élevée chez les individus Val/Val par rapport aux individus Met/Met [103]. Or, la dopamine est reconnue comme ayant un impact sur les fonctions exécutives des individus [104]. Un trouble des fonctions exécutives étant présent chez les patients Huntington [105], il est légitime de se demander si le polymorphisme Val^{158}Met impacte l'évolution de cette maladie.

Nous avons donc étudié l'évolution de la maladie en fonction du polymorphisme Val^{158}Met. Cette étude a pu être réalisée grâce au protocole « Biomarqueur » (*Predictive Biomarkers for Huntington's disease protocol*, NCT01412125).

Méthode

Cette étude a inclus 438 patients de la cohorte RHLF de 1994 à 2011. Tous ont signé un consentement, en accord avec le comité d'éthique de protection des personnes de l'hôpital Henri Mondor de Créteil. Seuls les patients hétérozygotes pour le gène IT15 (avec plus de 36 répétitions de CAG sur l'allèle muté) ont été inclus. Ils ont tous été génotypés pour le polymorphisme Val^{158}Met à l'hôpital de la Pitié Salpêtrière à Paris. Afin de comparer la distribution du polymorphisme à la population générale, 367 individus français ont été génotypés par la même technique.

A la première visite, 8% des patients étaient pré-symptomatiques, 39% étaient au stade I de la maladie, 30% au stade II, 18% au stade III et 5% au-delà. Ce large spectre permet de visualiser la progression de la maladie sur toute sa durée depuis l'apparition des premiers symptômes. Puis, ces patients ont été suivis annuellement, ce qui nous a permis de constituer une cohorte longitudinale de 406 patients. Ces patients ont tous été testés grâce aux échelles de l'UHDRS. Ainsi, nous pouvons mesurer l'impact du polymorphisme Val^{158}Met sur les troubles moteurs, fonctionnels et cognitifs.

Nous avons modélisé l'évolution des performances motrices, fonctionnelles et cognitives depuis le début de la maladie. La période d'observation étant longue, nous avons tenu compte des effets plafond et plancher associés aux tests de l'UHDRS grâce aux modèles à variable latente [106]. En effet, nous supposons que la variable modélisant la progression

de la maladie évolue linéairement avec le temps mais n'est pas directement observable (variable latente). Les scores obtenus grâce aux tests de l'UHDRS permettent d'évaluer la progression de la maladie et sont des transformations non linéaires (par exemple, transformation Beta) de la variable latente d'intérêt.

La variable latente d'intérêt est quant à elle expliquée par un modèle linéaire mixtes pour données longitudinales où les covariables sont le nombre de répétitions de CAG, le polymorphisme (Met/Met, Val/Val ou Met/Val) et le niveau d'études.

Discussion

Les résultats montrent que le polymorphisme Val^{158}Met joue un rôle dans la progression de la maladie sur le plan moteur et cognitif. Les patients homozygotes Met/Met ont des performances cognitives plus élevées que les patients homozygotes Val/Val au début de la maladie, mais leurs performances diminuent plus vite au cours du temps. Le polymorphisme Val^{158}Met apparaît donc comme un bon marqueur pronostique de l'évolution des troubles cognitifs dans la maladie de Huntington. La connaissance de ce marqueur a un double intérêt. Tout d'abord, dans la mise en place des essais cliniques, que nous discutons dans la section 5.2 suivante. Son second intérêt réside dans la mise en place d'un traitement « personnalisé » pour les patients Huntington. En effet, nous faisons l'hypothèse que les meilleures performances cognitives du groupe Met/Met au début de la maladie sont induites par une plus forte présence de dopamine dans le cortex préfrontal de ces patients par rapport aux patients Val/Val. Cependant, au fur et à mesure de la progression de la maladie, cet excès de dopamine pourrait être toxique, accélérant le processus d'atrophie [107]. Ici nous suggérons que les patients homozygotes Val/Val reçoivent un traitement par neuroleptiques au début de leur maladie. Les neuroleptiques permettent de diminuer le niveau de dopamine dans le striatum et de diminuer l'activité du gène COMT dans le cortex préfrontal.

Dans cette analyse, nous avons tenu compte des effets planchers et plafond des tests lorsque ceux-ci sont utilisés à des stades très précoces ou très tardifs de la maladie. L'utilisation de modèles linéaires mixtes sans transformation beta donne des résultats similaires avec un déclin cognitif plus rapide chez les patients Met/Met. Cependant, les analyses de sensibilité sur des sous-échantillons ne montrent pas de différences significatives selon le polymorphisme bien que la tendance reste la même. Cette étude va dans le sens d'un effet de la COMT sur le déclin cognitif des patients atteints de la maladie de Hutington, mais il serait nécessaire de le confirmer par une nouvelle étude intégrant plus de données associées aux premiers mois de la maladie. De plus, une nouvelle étude devrait intégrer les données relatives aux traitements. La maladie de Huntington est expliquée en grande partie par le nombre de répétitions CAG sur le gène HTT mais d'autres gènes, comme la COMT, peuvent avoir un effet additif sur le déclin.

COMT Val^{158}Met Polymorphism Modulates Huntington's Disease Progression

Short title: *COMT* Polymorphism in Huntington's Disease

Ruth de Diego-Balaguer[1,2,3,4*], Catherine Schramm[5,6,7*], Isabelle Rebeix[8,9], Emmanuel Dupoux[5,10], Alexandra Durr[8,11], Alexis Brice[8,11], Perrine Charles[11], Laurent Cleret de Langavant[5,6,7,12], Katia Youssov[5,6,7,12], Christophe Verny[13], Vincent Damotte[8,9], Jean-Philippe Azulay[14], Cyril Goizet[15], Clémence Simonin[16,17], Christine Tranchant[18], Patrick Maison[6,7,19], Amandine Rialland[19], David Schmitz[19], the French Speaking Huntington Group, Charlotte Jacquemot[5,6,7], Bertrand Fontaine[9,11], Anne-Catherine Bachoud-Lévi[5,6,7,12]

*These two authors contributed equally to the study

[1]ICREA, 08010 Barcelona, Spain

[2]Universitat de Barcelona, Departament de Psicologia Bàsica, 08035 Barcelona, Spain

[3]IDIBELL, Unitat de Cognició i Plasticitat Cerebral, 08907 L'Hospitalet de Llobregat, Spain

[4]Institut de Neurociència, Universitat de Barcelona

[5]Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 75005 Paris, France

[6]INSERM U955, Equipe 01 Neuropsychologie Interventionnelle, 94000 Créteil, France

[7]Université Paris Est, Faculté de Médecine, 94000 Créteil, France

[8]INSERM-UPMC-CNRS, UMR 7225-1127, Institut Cerveau Moelle-ICM, Hôpital Pitié-Salpêtrière, 74013 Paris, France

[9]Assistance Publique-Hôpitaux de Paris, Département des Maladies du Système Nerveux, Hôpital Pitié-Salpêtrière, 74013 Paris, France

[10]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-EHESS-CNRS, Paris, 75005, France

[11]Assistance Publique-Hôpitaux de Paris, Département de Génétique, Hôpital Pitié-Salpêtrière, 74013 Paris, France

[12]Assistance Publique-Hôpitaux de Paris, Centre de Référence Maladie de Huntington, Service de Neurologie, Hôpital Henri Mondor-Albert Chenevier, 94000 Créteil, France

[13]CHU d'Angers, Centre de Référence des Maladies Neurogénétiques, Service de Neurologie; 49933 Angers, France

[14]CHU de Marseille - Hôpital de la Timone, Service de Neurologie et Pathologie du Mouvement, 13385 Marseille, France

[15]CHU de Bordeaux-GH Sud - Hôpital Haut-Lévêque, Service de Neurologie, 33604, Pessac, France

[16]CHRU de Lille, Service de Neurologie et Pathologie du Mouvement, 59000, Lille, France

[17]INSERM UMR-S 1172, JPArc, centre de recherche Jean-Pierre-Aubert neurosciences et cancer, Université de Lille, 59000 Lille, France

[18]CHU de Strasbourg - Hôpital de Hautepierre, Service de Neurologie, 67098 Strasbourg, France

[19]Assistance Publique-Hôpitaux de Paris, Hôpital Henri Mondor, Unité de Recherche Clinique, 94000 Créteil, France

Corresponding author: Prof. AC Bachoud-Lévi

Centre de référence – maladie de Huntington

Unité de Neurologie Cognitive, Hôpital Henri Mondor

51 avenue du Maréchal de Lattre de Tassigny

94010 Créteil, France

bachoud@gmail.com

Tel +33 1 49 81 23 15; Fax +33 1 49 81 23 26

Abstract

Little is known about the genetic factors modulating the progression of Huntington's disease (HD). Dopamine levels are affected in HD and modulate executive functions, the main cognitive disorder of HD. We investigated whether the Val^{158}Met polymorphism of the *catechol-O-methyltransferase* (*COMT*) gene, which influences dopamine (DA) degradation, affects clinical progression in HD. We carried out a prospective longitudinal multicenter study from 1994 to 2011, on 438 HD gene carriers at different stages of the disease (34 pre-manifest; 172 stage 1; 130 stage 2; 80 stage 3; 17 stage 4; and 5 stage 5), according to Total Functional Capacity (TFC) score. We used the Unified Huntington's Disease Rating Scale to evaluate motor, cognitive, behavioral and functional decline. We genotyped participants for *COMT* polymorphism (107 Met-homozygous, 114 Val-homozygous and 217 heterozygous) and 367 controls of similar ancestry. We compared clinical progression, on each domain, between groups of *COMT* polymorphisms, using latent-class mixed models accounting for disease duration and number of CAG (cytosine adenine guanine) repeats. We show that HD gene carriers with fewer CAG repeats and with the Val allele in the COMT polymorphism displayed slower cognitive decline. The rate of cognitive decline was greater for Met/Met homozygotes, which displayed a better maintenance of cognitive capacity in earlier stages of the disease, but had a worse performance than Val allele carriers later on. The *COMT* polymorphism did not significantly impact functional and behavioral performance. Since *COMT* polymorphism influences progression in HD, it could be used for stratification in future clinical trials. Moreover, DA treatments based on the specific *COMT* polymorphism and adapted according to disease duration could potentially slow HD progression.

Abbreviations

CAG: Cytosine-Adenine-Guanine

COMT: Catechol-O-Methyltransferase

DA: Dopamine

FAS: Functional Assessment Scale

HD: Huntington's disease

Htt: Huntingtin

ICC: Intraclass Correlation Coefficient

IS: Independence Scale

Met: Methionine

mHtt: Mutant Huntingtin

PFC: Prefrontal Cortex

SD: Standard Deviation

SDMT: Symbol Digit Modalities Score

SE: Standard Error

TFC: Total Function Capacity

UHDRS: Unified Huntington's Disease Rating Scale

Val: Valine

INTRODUCTION

Huntington's disease (HD) is an autosomal dominant inherited neurodegenerative disease caused by increased number of CAG (cytosine adenine guanine) repeats in the Huntingtin (*Htt*) gene on chromosome 4 [1]. It primarily affects the striatum and manifests as progressive motor, behavioral and cognitive disturbances, leading to death about 15 to 20 years after onset. There is currently no effective course-modifying treatment.

Phenotypic expression differs considerably between patients. Age at onset varies and few of the underlying genetic factors have been identified [2]. The size of the number of CAG repeats in the mutated *Htt* (*mHtt*) gene is inversely related to age at onset of HD patients, but accounts for only 40 to 70% of its variance [3]. The implication of other genes in HD such as the PPARGC1A, GRIK2, APOE and BDNF genes, has been shown, but their impact was not replicated in subsequent studies [4, 5, 6]. The factors influencing disease progression remain to be identified [7]. Higher number of CAG repeats in the *mHtt* gene is associated with faster motor, cognitive, and functional decline [8]. The influence of the number of CAG repeats in the normal *Htt* allele remains uncertain, either on age at onset or disease progression [3, 9].

Here, in addition to results provided by genome wide association mapping conducted on the motor onset [10], we conduct an *a priori* study on the *catechol-O-methyltransferase* (*COMT*) to assess its impact on HD evolution [11]. Indeed, it is reasonable to hypothesize that *COMT* may play a role in HD. *COMT* degrades catecholamines, such as dopamine (DA). Medium-sized striatal spiny GABAergic neurons bearing dopaminergic receptors (D1 and D2) are preferentially affected in HD [12]. The density of these receptors in the striatum decreases [13] along with DA and GABA concentrations in HD patients. In the normal population, a

valine-to-methionine substitution in position 158 ($Val^{158}Met$) on the *COMT* gene on chromosome 22 increases *COMT* activity, to levels 38% higher for the Val/Val genotype than for the Met/Met genotype [14], resulting in lower DA levels in Val/Val patients. *COMT* polymorphism essentially affects DA levels in the prefrontal cortex (PFC) because striatal DA levels are regulated principally by the DA transporter (DAT). However, *COMT* polymorphism influences the severity of cognitive and behavioral symptoms in other diseases affecting subcortical DA regulation, such as Parkinson's disease [15, 16] and schizophrenia [17], and is predictive of disease progression and psychosis in 22q11.2 deletion syndrome [18]. In HD, *COMT* polymorphism has no influence on motor onset [4], but its effect in behavioral, cognitive and functional domains and in disease progression remains to be investigated. The cognitive effects of the *COMT* polymorphism in various diseases and in the healthy population have repeatedly been reported to be specific to executive functions (see [19, 20] for reviews), and executive function defects are the hallmark cognitive dysfunction in HD. Furthermore, even at low doses, DA aggravates *mHtt* toxicity in striatal neuron cultures [21] and increases behavioral and motor deficits in YAC128 mice [22], a transgenic model of HD. Thus, *COMT* polymorphism may affect the progression of HD.

We investigated the impact of *COMT* polymorphism on HD progression in a longitudinal prospective study, and found that it affects cognitive and motor declines but has no impact on behavioral and functional declines.

MATERIAL AND METHODS

Participants

We report a longitudinal prospective long-term study of 438 HD gene carriers from the Predictive Biomarkers for Huntington's disease protocol (NCT01412125), which was approved by the ethics committee of Henri Mondor Hospital (Créteil, France) in accordance with EU and French bioethics laws. All HD gene carriers gave written informed consent. They were heterozygous for the *Htt* gene (> 36 CAG repeats in *mHtt*) and aware of their genetic status. They had no other neurological conditions or long-term experimental treatment (e.g. cell transplantation).

Data were collected from 1996 to 2011, at eight centers from the French Speaking Huntington's Disease Group (Angers: 24%, Bordeaux: 7%, Créteil: 34%, Lille: 4%, Lyon: 1%, Marseille: 12%, Paris: 11%, Strasbourg: 7%), and centralized at the National Reference Centre for Huntington's disease in Créteil. The date at onset has been available for 86.53% of the HD gene carriers. It corresponds to the apparition of first symptoms and it was determined (observed) by the clinician (93.14%) or, if missing, by the family (5.28%), or, if missing too, by the participant (1.58%).

Blood samples were centralized at the DNA bank of Pitié-Salpêtrière Hospital. The number of CAG repeats was routinely determined [23]. The rs4680 (*COMT* Val^{158}Met) polymorphism was genotyped by PCR with appropriate primers [24]. We investigated the distribution of *COMT* genotypes in the general population, by genotyping 367 independent French controls with the same technique.

Clinical assessment

HD gene carriers were followed up with the Unified Huntington's Disease Rating Scale (UHDRS) [25], which combines motor, functional, behavioral and cognitive assessments. Motor domain was assessed using the Total Motor Score (TMS, range: 0 to 124). Functional domain was assessed using the Total Functional Capacity scale (TFC, range: 13 to 0), Functional Assessment Scale (FAS, range: 25 to 50) and Independence Scale (IS, range: 100 to 0). Behavioral domain was assessed using the psychiatric part of the UHDRS (range: 0 to 88). Cognitive domain was assessed using the Stroop Test (color naming: Stroop C, word reading: Stroop W, and color-word interference: Stroop C/W), Symbol Digit Modality Test (SDMT), and letter fluency (for P, R and V in French). For letter fluency, testing at two minutes appears to be more sensitive than testing at one minute [26]. The French version used in this study includes both measurements. Higher scores in IS, FAS and TMS indicate greater impairment. For all other tasks, higher scores indicate lower impairment.

The first evaluation corresponding to the entrance in the study (first visit) occurred before onset (pre-manifest) in some individuals and at various times after onset in others, such that the sample encompassed the entire spectrum of HD progression (first visit: 8% pre-manifest gene carriers; 39% patients at Stage 1; 30% Stage 2; 18% Stage 3; 4% stage 4; and 1% Stage 5). Pre-manifest gene carriers were defined by as having a TMS below or equal 5 [27], and a TFC score of 13. The visits were performed annually, with few exceptions, with a mean inter-visit delay of 1.2 years (SD = 0.4). The mean number of visits per HD gene carriers was 5.0 (SD = 3.2; range: 1 to 19 visits). Thirty-two HD gene carriers were seen only once. Data were recorded for 2185 visits. The mean duration of follow-up was 4.3 years (SD = 3.0; range: 0 to 15.5 years).

Statistical analyses

Demographics and characteristics of the COMT polymorphism groups at the first visit

The χ^2 goodness-of-fit test was computed to compare the distribution of *COMT* genotypes in HD gene carriers and in the control group.

We assessed whether baseline characteristics of HD gene carriers were similar in the different *COMT* polymorphism samples (Met/Met, Val/Val and Met/Val), by first assessing the differences between groups for each score of the UHDRS at the first visit. Demographic data and clinical characteristics of the sample (N = 438) at the first visit were compared between groups, with a Pearson's χ^2 tests for qualitative variables and a one-way ANOVA for quantitative variables. For variables with significant difference between groups, student's *t*-tests (or Welch's tests in cases of unequal variances) were performed with Bonferroni correction for multiple pairwise comparisons (see supplemental data S1 Table for the same comparisons in the subgroup included in the longitudinal analysis).

Number of CAG repeats and age at onset

We first assessed the impact of the number of CAG repeats on age at onset. We used a linear regression model, with age at onset as the dependent variable and the number of CAG repeats as an independent variable. The R^2 value provided by the model is an estimate of the proportion of the variability of the age at onset explained by the number CAG repeats.

We also calculated an expected age at onset according to the Langbehn et al. model [28], derived from the number of CAG repeats using the formula: *expected age = (21.54 + exp (9.556 - 0.146*CAG))*. We evaluated the agreement between this expected age at onset and the age at onset provided in our database by calculating the intraclass correlation coefficient

(ICC), a measure for concordance. The ICC was obtained by a two-way mixed effect model [29].

Longitudinal analysis of disease progression

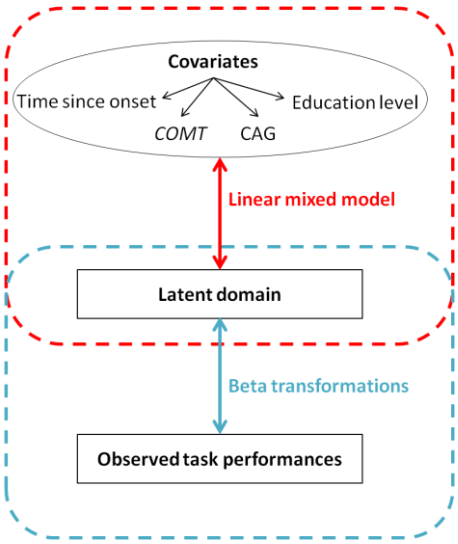
The longitudinal analysis was conducted in HD gene carriers assessed at least twice and for which the date of onset was known, and included 350 HD gene carriers and 1912 visits.

We compared progression over time between groups, by calculating the overall change in motor, functional, behavioral and cognitive domains per year since the date of onset. Domains are not observable *per se* but are modeled by a latent variable reflected by observed performances at each task. We performed four latent-class mixed models [30], one per domain, where each model combines (i) a linear mixed model to explain latent domain according to covariates and (ii) beta transformations which link observed performances at each task to latent domain (Fig 1). Similarly to classical linear mixed models, the latent-class mixed model allows integrating data from HD gene carriers with unequal duration of follow-up and introducing a subject-specific intercept by random effects to account for within-unit correlation for outcome and between-subject variability [31]. These models take into account all observations for each patient, without listwise deletion. Moreover, the use of beta transformations allows taking into account the ceiling and floor effects of UHDRS tasks.

Parameters of linear mixed model and beta transformations are estimated simultaneously using maximum likelihood method and Monte-Carlo integration. The disease duration (time since onset), *COMT* polymorphism, number of CAG repeats in *mHtt*, interaction between disease duration and the number of CAG repeats, interaction between disease duration and *COMT* polymorphism and education level were retained as covariates. For *COMT* polymorphism, included in the model as a categorical covariate, Met/Met was the reference

group, allowing the comparison between Met/Met and Val/Val genotypes and between Met/Met and Met/Val genotypes. We compared Met/Val and Val/Val genotypes by recomputing the models with Met/Val genotype as the reference group. All *P*-values were adjusted with Bonferroni correction in two steps: one within the *COMT* polymorphism groups comparison and one for multiple comparisons across domains. Based on Akaike's information criterion and Bayesian information criterion [32], number of CAG repeats in the normal *Htt* allele and CAG-*COMT* interaction did not improve model fit, thus they were removed from the final model.

Fig 1: Structure of the latent class mixed models



To assess the robustness of the results, a sensitivity analysis was performed excluding outliers, on the basis of the number of CAG repeats and of the distribution of dates of visits in our cohort (see supplemental data S1 Figure, S2 Figure). The sensitivity analysis included HD gene carriers with a number of CAG repeats between 39 and 49 that were followed in the 20 years after disease onset.

Analyses were conducted with R 2.3 software (http://www.r-project.org/). The R package lcmm was used to perform the longitudinal analysis. All tests were two-tailed. Values of $P < 0.05$ were considered significant.

RESULTS

Demographics and characteristics of the COMT polymorphism groups at the first visit

The χ^2 goodness-of-fit test confirms that the distribution of *COMT* genotypes is similar in HD gene carriers and the control group ($P = 0.15$) (see Table 1).

Table 1. Distribution of the *COMT* genotypes in HD gene carriers and control groups

	Met/Met	Met/Val	Val/Val
Controls *N* (%)	70 (19.1)	202 (55.0)	95 (25.9)
HD gene carriers *N* (%)	107 (24.4)	217 (49.6)	114 (26.0)

HD: Huntington's disease; Met: Methionine; Val: Valine

Demographic and clinical data of HD gene carriers for the first visit are displayed in Table 2. Baseline demographic and clinical characteristics are similar for all *COMT* polymorphisms (one-way ANOVA, $P > 0.05$) except that HD gene carriers with the Met/Val genotype have a lower educational level than those with the Val/Val (pairwise comparison, corrected $P = 0.01$) or Met/Met (pairwise comparison, corrected $P = 0.01$) genotypes. (See supplemental data S1 Table for descriptive analysis of HD gene carriers included in the longitudinal analysis.)

280 **Table 2. Demographic characteristics and performance of HD gene carriers**
281

	Met/Met N=107	Met/Val N=217	Val/Val N=114	*P**
Age (yrs)	46.1 (12.8)	49.5 (12.1)	47.9 (11.2)	Ns
Sex (% men)	55.1	47.0	52.6	Ns
Age at onset (yrs)	41.9 (11.6)	45.3 (11.5)	43.6 (9.7)	Ns
Educational level (yrs in education)	12.3 (3.4)	11.2 (2.9)	12.2 (3.3)	0.0012
BMI	22.6 (3.7)	22.7 (3.6)	22.1 (3.5)	Ns
CAG repeats *mHtt*	45.3 (4.5)	44.5 (3.6)	44.6 (3.1)	Ns
CAG repeats *Htt*	18.3 (2.8)	18.9 (4.1)	18.9 (3.9)	Ns
Antipsychotic use (%)	75.7	73.3	72.8	Ns
Antidepressant use (%)	28.0	27.6	28.1	Ns
Benzodiazepine use (%)	24.3	23.0	14.0	Ns
UHDRS				
TMS	30.6 (19.7)	32.3 (22.0)	35.9 (23.3)	Ns
Behavior	17.8 (13.4)	17.1 (11.0)	16.3 (12.0)	Ns
FAS	29.4 (5.3)	30.0 (5.9)	30.7 (6.0)	Ns
IS	84.3 (15.2)	83.0 (16.7)	81.1 (16.4)	Ns
TFC	9.4 (3.4)	9.3 (3.4)	8.8 (3.6)	Ns
L Fluency 1’	22.7 (12.9)	20.0 (12.6)	19.7 (13.1)	Ns
L Fluency 2’	33.3 (21.4)	28.8 (19.9)	28.2 (20.6)	Ns
Stroop W	64.4 (24.9)	61.9 (23.3)	65.4 (27.5)	Ns
Stroop C	47.8 (20.6)	43.8 (17.1)	46.7 (20.7)	Ns
Stroop W/C	25.8 (14.3)	23.6 (12.8)	23.6 (15.4)	Ns
SDMT	26.6 (16.6)	24.1 (15.2)	25.1 (17.3)	Ns

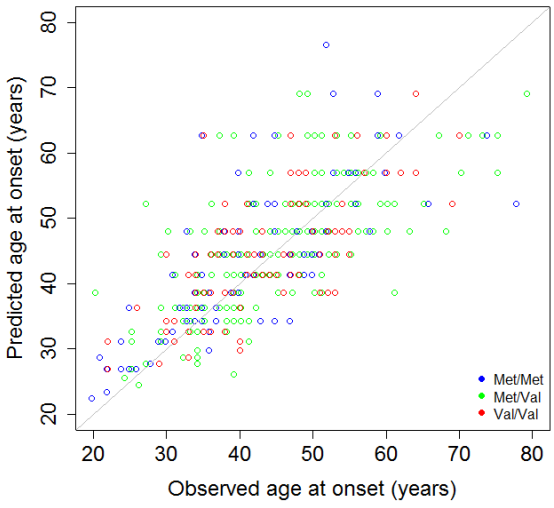
HD: Huntington's disease; BMI: body mass index; CAG repeats refers to the number of CAGs in the mutated (*mHtt*) and non-mutated (normal *Htt*) alleles of the Huntingtin gene; UHDRS: Unified Huntington's Disease Rating Scale; TMS: Total Motor Score; FAS: Functional Assessment Scale; IS: Independence Scale; TFC: Total Functional Capacity; Letter fluency (L Fluency) was tested with PRV letters (French norms) at 1 minute (1') and 2 minutes (2'); Stroop C: Color; W: Word; W/C: Word/Color (interference score); SDMT: Symbol Digit Modalities Test. Quantitative variables are presented as means, with the standard deviation in brackets, and qualitative variables are presented as frequency counts. Medication use is expressed as a percentage.*Non corrected *P*-values; Chi-squared test for qualitative variables and one-way ANOVA for quantitative data; Ns: not significant.

Number of CAG repeats and age at onset

The number of CAG repeats explains 49.61% of the variability of age at onset (β coefficient = -2.07 (SE = 0.11), $P < 0.001$).

The ICC measuring agreement between expected age at onset by formula (1) and age at onset provided in the database is high for the whole cohort (0.71: [95% CI 0.65–0.76], $P < 0.0001$) and in each *COMT* group (Met/Met ICC = 0.75 [95% CI 0.64–0.83], $P < 0.0001$, Met/Val ICC = 0.70 [95% CI 0.62–0.77], $P < 0.0001$ and Val/Val ICC = 0.66 [95% CI: 0.54–0.76], $P < 0.0001$) (Fig 2).

Fig 2. Concordance between predicted and real age at onset.



Each point represents an individual patient. The observed age at onset is the one provided in the database. The predicted age at onset is the one calculated by the formula $21.54 + \exp(9.556 - 0.146 \times CAG)$. The gray line is the first bisector corresponding to the line of predicted=observed. The closeness of the points to the gray line indicates the extent to which predicted age at onset matches real age at onset. If predicted age at onset is greater than the observed age at onset, the points are located above the gray line. By contrast, if the predicted age at onset is below the real age at onset, the points are located below the gray line.

Longitudinal analysis of disease progression

Table 3 displays the modeling parameters of the linear mixed models corresponding to the disease evolution within the four domains: motor, behavior, functional and cognitive. After correcting *P*-values, there is no effect of *COMT* polymorphism or the number of CAG repeats on latent processes at time 0 (estimated onset). A higher education level is correlated with higher performance in cognitive and functional domains. For all *COMT* polymorphism, performance declined over time for the motor, cognitive, and functional domains but not for behavior (see Fig 3). Higher number of CAG repeats is associated with a faster decline for motor, cognitive and functional domains. Met/Met HD gene carriers decline faster than Val/Val and Met/Val HD gene carriers in cognitive domain. Met/Val HD gene carriers decline faster than Val/Val HD gene carriers in motor domain. At age at onset and over the 10 years following disease onset, Met/Met HD gene carriers outperform Met/Val and Val/Val HD gene carriers in the cognitive domain. However, since they decline faster they subsequently perform less well than the other HD gene carriers (Fig 3 and Fig 4). The intersection of the progression curves for the Met/Met and Met/Val groups is estimated at 7.2 years for the cognitive domain. The intersection of the Met/Met and Val/Val curves is estimated at 10.9 years for the cognitive domain. The intersection of the Met/Val and Val/Val curves is estimated at 11.0 years for the motor domain.

Table 3. Impact of *COMT* genotype and the number of the number of CAG repeats in the long allele on disease evolution within the four domains

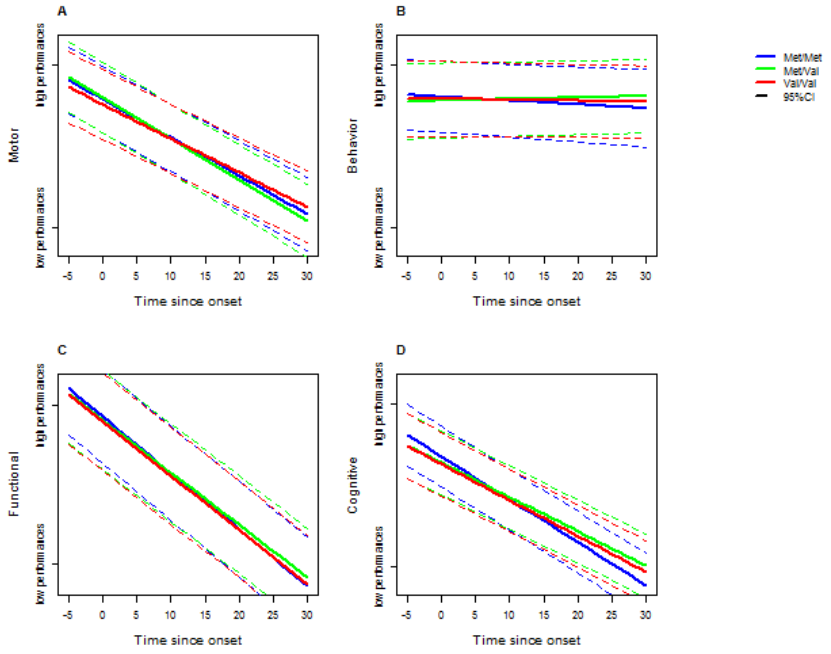
Domains	Motor (N=348)		Behavior (N=348)		Functional (N=348)		Cognitive (N=344)	
	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)
Baseline:								
Met/Val vs Met/Met	0.08 (0.18)	0.6673 (ns)	-0.46 (0.27)	0.0949 (ns)	-0.15 (0.17)	0.3745 (ns)	-0.30 (0.15)	0.0504 (ns)
Val/Val vs Met/Met	-0.26 (0.20)	0.1937 (ns)	-0.24 (0.30)	0.4186 (ns)	-0.22 (0.19)	0.2451 (ns)	-0.37 (0.17)	0.0337* (ns)
Val/Val vs Met/Val	-0.33 (0.16)	0.0434* (ns)	0.22 (0.23)	0.3418 (ns)	-0.07 (0.16)	0.6529 (ns)	-0.06 (0.15)	0.6580 (ns)
Number of CAG repeats	-0.01 (0.02)	0.4652 (ns)	0.07 (0.05)	0.1662 (ns)	0.04 (0.02)	0.0579 (ns)	0.03 (0.02)	0.0368* (ns)
Education level	0.03 (0.02)	0.0776 (ns)	0.05 (0.02)	0.0254* (ns)	0.05 (0.02)	0.0033** (0.0132*)	0.07 (0.02)	<0.0001*** (0.0001***)
Slope:								
Met/Val vs Met/Met	-0.01 (0.01)	0.2732 (ns)	0.06 (0.02)	0.0086** (ns)	0.02 (0.01)	0.1092 (ns)	0.04 (0.01)	<0.0001*** (<0.0001***)
Val/Val vs Met/Met	0.02 (0.01)	0.1535 (ns)	0.03 (0.03)	0.2102 (ns)	0.01 (0.01)	0.4185 (ns)	0.03 (0.01)	0.0002*** (0.0012**)
Val/Val vs Met/Val	0.03 (0.01)	0.0040** (0.0240*)	-0.03 (0.02)	0.1674 (ns)	-0.01 (0.01)	0.4714 (ns)	-0.01 (0.01)	0.2618 (ns)
Number of CAG repeats	-0.01 (0.001)	<0.0001*** (<0.0001)	-0.01 (0.004)	0.1250 (ns)	-0.01 (0.001)	<0.0001*** (<0.0001***)	-0.01 (0.001)	<0.0001*** (<0.0001***)

The motor domain was modeled including the performances at TMS; the behavioral domain was modeled including the performances at behavior task of the UHDRS; the functional domain was modeled including the performances at FAS and IS (TFC could not be included because there are not enough values for the model to converge); the cognitive domain was modeled including performances at letter fluency assessed at 1 and 2 minutes, SDMT and the three parts of the Stroop.

N: Number of HD gene carriers who have contributed to the estimation (cognitive tasks were not available for all HD gene carriers); SE: Standard error of the estimate, *P*: *P*-values (*** P<0.001, ** P<0.01, *P<0.05).

Baseline values correspond to the impact of covariates at estimated age at onset. *Slope* values correspond to the impact of covariates on the slope of the decline.

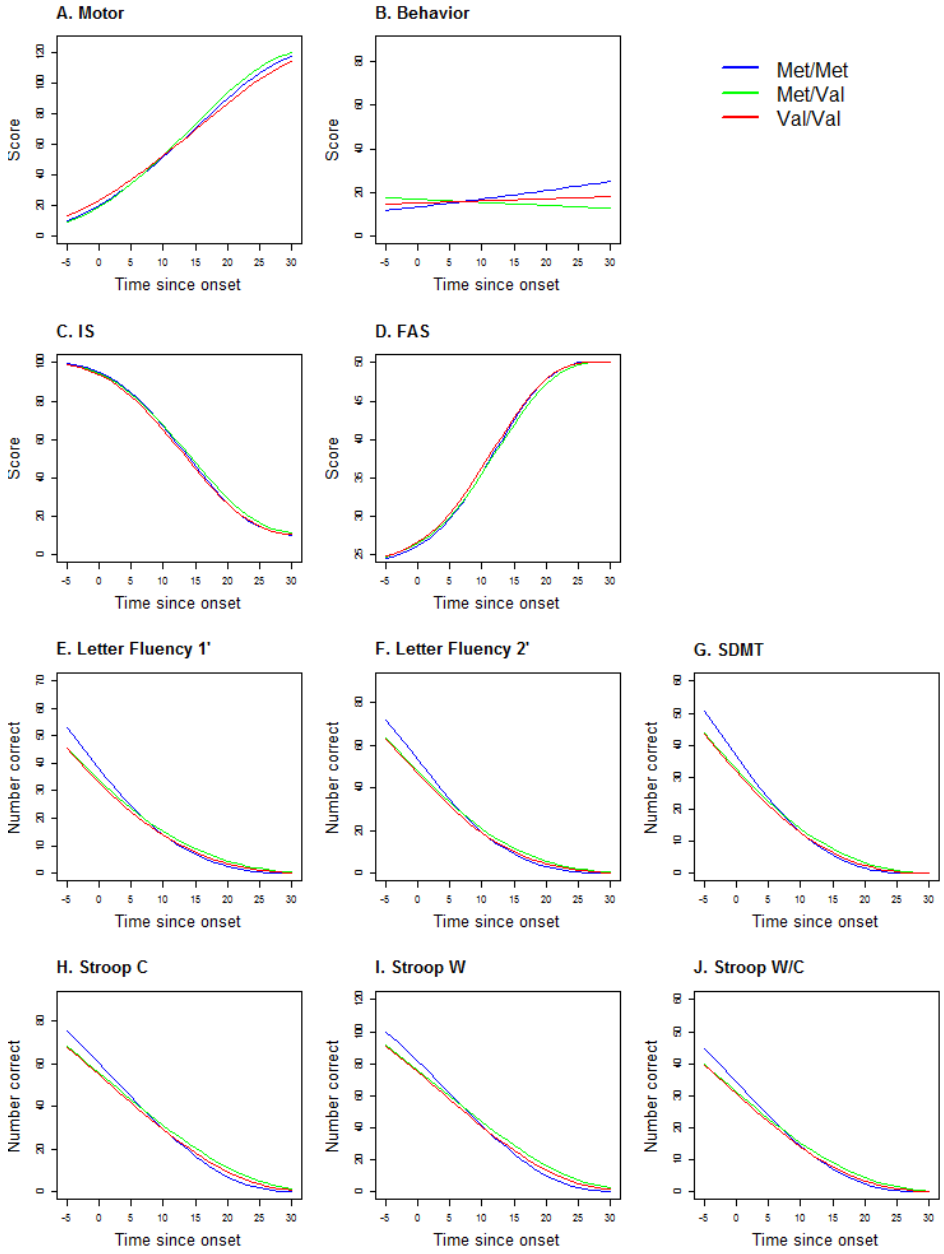
Fig 3. Curves of the impact of *COMT* polymorphism on the motor, behavioral, functional and cognitive domains, in a modeled cohort of HD patients with 45 CAG repeats and 12-year education level.



We plotted the evolution of performance as a function of time for each task. Performance decrease was represented by a negative slope. 45 CAG repeats is the mean number in the cohort studied. The latent motor process was modeled using the UHDRS motor score; the latent behavioral process was modeled using the UHDRS behavioral score; the latent functional process was modeled using the FAS and IS scores; The latent cognitive process was modeled using letter fluency at 1 minute, letter fluency at 2 minutes, SDMT, Stroop Color, Stroop Word and Stroop Word/Color interference.

The Figure 4 shows that for most tasks, the fit of disease is linear only for the first 15 years, displaying a floor effect after that point. Beta link functions between performance at each task and latent variable modeling of the domains are displayed on supplemental data S3 Figure.

Fig 4. Curves of the impact of *COMT* polymorphism on each UHDRS score, in a modeled cohort of HD patients with 45 CAG repeats and 12-year education level.



We plotted the evolution of performance for each task. 45 CAG repeats is the mean number in the cohort studied. UHDRS motor score **(A)**; UHDRS behavioral **(B)**, IS: Independence Score **(C)**; FAS: Functional Assessment Scale **(D)**, cognitive (letter fluency 1': at 1 minute **(E)**; letter fluency 2': at 2 minutes **(F)**; SDMT: symbol digit modalities test **(G)**; Stroop C: Stroop color **(H)**; Stroop W: Stroop word **(I)**; Stroop W/C: Stroop interference **(J)**.

In the sensitivity analysis, performance decline over time and larger number of CAG repeats are associated with a faster decline, in all domains except behavior. Met/Val HD gene carriers decline faster than Val/Val HD gene carriers in motor domain. Met/Met HD gene carriers

decline faster than Val/Val and Met/Val HD gene carriers in cognitive domain but the associated *P*-value is no longer significant after Bonferroni correction (see supplemental data S2 Table).

DISCUSSION

We investigated the impact of *COMT* polymorphism in a prospective multicenter study of 438 HD gene carriers all stages of HD followed up once yearly, during 4.3 (SD = 3.0) years, with the UHDRS. The *COMT* polymorphism distribution in this sample is similar to that reported for the European population [33]. As previous reported, the number of CAG repeats affects the age at onset and the disease progression in our cohort [3, 9]. Higher educational level improves cognitive performance at baseline, as observed in elder population [34]. The *COMT* polymorphism influences disease progression in cognitive domain in a biphasic manner. Met/Met HD gene carriers outperform Val/Val HD gene carriers in the cognitive domain during the first 10 years after disease onset. However, they then performed worse than Val/Val HD gene carriers, since their slope of decline is steeper. The effect of *COMT* polymorphism on motor domain is of particular interest because it modifies progression of motor performances rather than age at onset [4].

The *COMT* polymorphism does not influence disease progression in behavioral and functional domains.

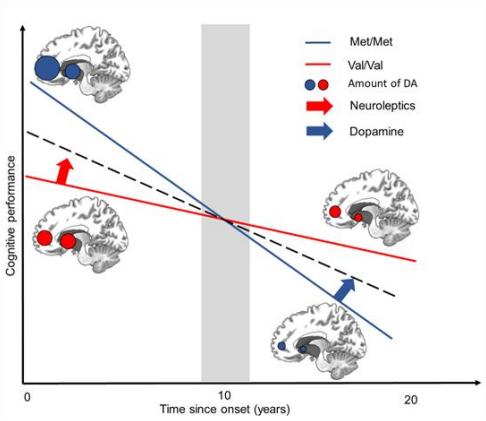
This study replicates the effect of the number of CAG repeats observed in other studies [35, 36]. It allows deciphering the effect of the *COMT* polymorphism presumably because unlike previous studies on other cohorts [37], we did not select HD gene carriers at particular disease stages or with specific number of CAG repeats. In addition we improved the value of our results by selecting the number of CAG repeats without including the age at onset as covariate

402 despite its known value [8] to avoid redundancy [38] since the age at onset and the number of
403 CAG repeats are two correlated factors [3, 28]. Furthermore, the use of a single language for
404 cognitive testing decreased inter-subject variability in cognitive performance. The HD gene
405 carriers were followed up prospectively for as long as possible, from pre-manifest to
406 advanced stages. Although most of the data was collected between 5 and 15 years after
407 disease onset, it provides a unique continuum of disease progression with enough follow up
408 data to conduct a longitudinal analysis.
409 The latent-class mixed model has the advantage of grouping several tasks within domains and
410 provides a global picture by domain without focusing on specific tasks. This approach,
411 recently developed, is already used in studies evaluating the cognitive decline [39, 40]. To
412 ensure that it models disease progression as well as the classical task by task multiple linear
413 mixed model [41, 42], we ran both latent-class mixed models and the linear mixed models on
414 our data (see supplemental data S3 Table). Both models show higher cognitive decline for the
415 Met/Met group. The latent-mixed model has also the advantage to avoid the calculation of
416 sum of performance for tasks with different weights and to take into account all assessments
417 and not only a delta between baseline and last assessment as in some regression linear
418 analyses [43].
419 Our study shows that the impact of the *COMT* polymorphism differs according to each
420 domain like in previous studies of Parkinson's disease and schizophrenia [17, 18, 44]. The
421 cognitive assessment in this study evaluated executive functions, the principal functions
422 affected in HD. These functions are modulated by the *COMT* polymorphism, improving with
423 increases in DA availability in healthy individuals with the Met/Met genotype [45]. These
424 effects on disease progression have implications for our understanding of the dynamics of DA
425 in the PFC and striatum in HD. *COMT* influences DA levels, mostly in the PFC, consistent
426 with the specific effect on cognitive symptoms observed in HD. Indeed, DA antagonists with

systemic action, which reduce DA levels in both the PFC and striatum, have been shown to worsen cognitive impairment [22] and chorea intensity at early stages. As in healthy individuals [46], the higher availability of DA in the Met/Met genotype is associated with a preservation of cognitive function at early stages. The greater availability of DA in Met/Met individuals appears to have an effect similar to cognitive reserve in the initial stages of the disease. The effects of high DA levels, which are initially beneficial in the early stages of the disease, eventually become detrimental, due to the long-term toxicity of DA in striatal cells [21].

This biphasic pattern over time suggests a symptomatic, rather than neuroprotective effect. Consistently, early and chronic treatment with the D2 antagonist haloperidol decanoate protects against neuronal dysfunction and aggregate formation in a rat model of HD [47]. *COMT* polymorphism also determines the response to entacapone [24] but not to levodopa. However, we cannot rule out the possibility of the Val allele being neuroprotective *per se*, because Met/Met individuals display greater gray matter degeneration within DA-innervated structures, including the striatum [45].

Fig 5. Schematic representation of the biphasic effect of *COMT* polymorphism in HD.



In the prefrontal cortex, DA levels are higher in Met/Met HD gene carriers at early stages and in HD gene carriers with premanifest disease than in controls. These levels subsequently decrease over time in both the Met/Met (in blue) and Val/Val groups (in red) [48]. The high levels of DA present in the PFC at early stages result in better cognitive performances. At late

stages, higher levels of DA in the PFC in Met/Met HD gene carriers may be toxic, increasing atrophy [21, 45]. In both *COMT* polymorphis groups, the level of striatal DA decreases over time.

These results open up new possibilities for treatments tailored to patient genotype, slowing disease progression, especially for treatments controlling cognitive function decline, which are currently lacking. It should pave the way for personalized treatment in HD gene carriers by adapting treatment to time- and region-specific changes, taking *COMT* genotype into account. At early stages of the disease, the combination of treatments decreasing DA levels in the striatum and *COMT* inhibitors increasing DA levels in the PFC, might prevent the exacerbation of cognitive deficits, or even improve cognitive ability (Fig 5) in Val/Val HD gene carriers. It has an immediate application in pharmacological management of HD, as inhibitors or activators of *COMT* are already available. At later stages, more than 10 years after onset, it may be harder to target DA levels in the PFC specifically, as classical antipsychotic drugs occupy a large proportion of subcortical dopamine D2 receptors, whereas atypical antipsychotics preferentially occupy cortical 5-HT(2) receptors.

Our study also has practical implications for future clinical trials assessing decline in HD because *COMT* polymorphism appears as an important factor of stratification. Moreover, the methodology we used could be adapted to other neurodegenerative diseases.

Acknowledgments

We thank C. Lalanne for comments relating to methodology and Page Piccinini for her language corrections.

References

1. The Huntington Collaborative Study Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72: 971-983.

2. Wexler NS. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci USA. 2004;10: 3498-3503.

3. Lee J-M, Ramos EM, Lee J-H, Gillis T, Mysore JS, Hayden MR, et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology. 2012;78: 690-695. doi:10.1212/WNL.0b013e318249f683

4. Ramos EM, Latourelle JC, Gillis T, Mysore JS, Squitieri F, Di Pardo A, et al. Candidate glutamatergic and dopaminergic pathway gene variants do not influence Huntington's disease motor onset. Neurogenetics. 2013;14: 173-179. doi:10.1007/s10048-013-0364-y

5. Arning L, Epplen JT. Genetic modifiers in Huntington's disease: fiction or fact? Neurogenetics. 2013;14: 171-172. doi:10.1007/s10048-013-0365-x

6. Gusella JF, MacDonald ME, Lee J-M. Genetic modifiers of Huntington's disease. Mov Disord. 2014;29: 1359-1365. doi:10.1002/mds.26001

7. Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol. Elsevier Ltd; 2011;10: 83-98. doi:10.1016/S1474-4422(10)70245-3

8. Rosenblatt A, Kumar B V, Mo A, Welsh CS, Margolis RL, Ross C A. Age, CAG repeat length, and clinical progression in Huntington's disease. Mov Disord. 2012;27: 272-276. doi:10.1002/mds.24024

9. Aziz NA, Jurgens CK, Landwehrmeyer GB, Van Roon-Mom WMC, Van Ommen GJB, Stijnen T, et al. Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. Neurology. 2009;73: 1280-1285. doi:10.1212/WNL.0b013e3181bd1121

10. Lee JM, Wheeler VC, Chao MJ, Vonsattel JPG, Pinto RM, Lucente D, et al. Identification of genetic factors that modify clinical onset of Huntington's disease. Cell. 2015;162(3): 516-526.

11. Bečanović K, Nørremølle A, Neal SJ, Kay C, Collins JA, Arenillas D, et al. A SNP in the HTT promoter alters NF-[kappa] B binding and is a bidirectional genetic modifier of Huntington disease. Nature neuroscience. 2015;18(6), 807-816.

12. Vonsattel JP, DiFiglia M. Huntington disease. J Neuropathol Exp Neurol. 1998;57: 369-384.

13. Turjanski N, Weeks R, Dolan R, Harding AE, Brooks DJ. Striatal D 1 and D 2 receptor binding in patients with Huntington's disease and other choreas A PET study. Brain. 1995;118: 689-696. doi:10.1093/brain/118.3.689

14. Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S, et al. Functional analysis of genetic variation in catechol-O-methyltransferase (COMT): effects on mRNA, protein, and enzyme activity in postmortem human brain. Am J Hum Genet. 2004;75: 807-821. doi:10.1086/425589

15. Foltynie T, Brayne CEG, Robbins TW, Barker RA. The cognitive ability of an incident cohort of Parkinson's patients in the UK. The CamPaIGN study. Brain. 2004;127: 550-560. doi:10.1093/brain/awh067

16. Wu K, O'Keeffe D, Politis M, O'Keeffe GC, Robbins TW, Bose SK, et al. The catechol-O-methyltransferase Val(158)Met polymorphism modulates fronto-cortical dopamine turnover in early Parkinson's disease: a PET study. Brain. 2012;135: 2449-2457. doi:10.1093/brain/aws157

17. Meyer-Lindenberg A, Miletich RS, Kohn PD, Esposito G, Carson RE, Quarantelli M, et al. Reduced prefrontal activity predicts exaggerated striatal dopaminergic function in schizophrenia. Nat Neurosci. 2002;5: 267-271. doi:10.1038/nn804

18. Gothelf D, Eliez S, Thompson T, Hinard C, Penniman L, Feinstein C, et al. COMT genotype predicts longitudinal cognitive decline and psychosis in 22q11.2 deletion syndrome. Nat Neurosci. Nature Publishing Group; 2005;8: 1500-1502. doi:10.1038/nn1572

19. Cools R, D'Esposito M. Inverted-U-shaped dopamine actions on human working memory and cognitive control. Biol Psychiatry. Elsevier Inc.; 2011;69: e113-125. doi:10.1016/j.biopsych.2011.03.028

20. Green AE, Munafò MR, DeYoung CG, Fossella JA, Fan J, Gray JR. Using genetic data in cognitive neuroscience: from growing pains to genuine insights. Nat Rev Neurosci. Nature Publishing Group; 2008;9: 710-720. doi:10.1038/nrn2461

21. Charvin D, Vanhoutte P. Unraveling a role for dopamine in Huntington's disease: the dual role of reactive oxygen species and D2 receptor stimulation. Proc Natl Acad Sci USA. 2005;102(34), 12218-12223.

22. Tang T-S, Chen X, Liu J, Bezprozvanny I. Dopaminergic signaling and striatal neurodegeneration in Huntington's disease. J Neurosci. 2007;27: 7899-7910. doi:10.1523/JNEUROSCI.1396-07.2007

23. Harbo HF, Finsterer J, Baets J, Van Broeckhoven C, Di Donato S, Fontaine B, et al. EFNS guidelines on the molecular diagnosis of neurogenetic disorders: general issues, Huntington's disease, Parkinson's disease and dystonias. Eur J Neurol. 2009;16: 777-785. doi:10.1111/j.1468-1331.2009.02646.x

24. Corvol J-C, Bonnet C, Charbonnier-Beaupel F, Bonnet A-M, Fiévet M-H, Bellanger A, et al. The COMT Val158Met polymorphism affects the response to entacapone in Parkinson's disease: a randomized crossover clinical trial. Ann Neurol. 2011;69: 111-118. doi:10.1002/ana.22155

25. Kremer HPH, Hungtington Study Group X. Unified Huntington's disease rating scale: reliability and consistency. Movement Disorders. 1996;11: 136-142.

26. Cardebat D, Doyon B, Puel M, Goulet P, Joanette Y. Evocation lexicale formelle et sémantique chez des sujets normaux. Performances et dynamiques de production en fonction du sexe, de l'âge et du niveau d'étude. Acta Neurol Belg. Acta medica belgica; 90: 207-217.

27. Tabrizi, SJ, Langbehn, DR, Leavitt, BR, Roos, RA, Durr, A, Craufurd, D, et al. Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. The Lancet Neurology. 2009; 8(9): 791-801.

28. Langbehn DR, Hayden MR, Paulsen JS. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. Am J Med Genet B Neuropsychiatr Genet. 2010;153B: 397-408. doi:10.1002/ajmg.b.30992

29. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86: 420-428.

30. Proust-Lima C, Amieva H, Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. British Journal of Mathematical and Statistical Psychology. 2013;66(3): 470-487

31. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;4: 963-974.

32. Bozdogan H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika. 1987;52: 345-370. doi:10.1007/BF02294361

33. Palmatier MA, Kang AM, Kidd KK. Global variation in the frequencies of functionally different catechol-O-methyltransferase alleles. Biol Psychiatry. 1999;46: 557-567. doi:10.1016/S0006-3223(99)00098-0

34. Wilson RS, Hebert LE, Scherr PA, Barnes LL, De Leon CM, Evans DA. Educational attainment and cognitive decline in old age. Neurology. 2009;72(5): 460-465.

35. Kieburtz K, MacDonald M, Shih C, Feigin A, Steinberg K, Bordwell K, et al. Trinucleotide repeat length and progression of illness in Huntington's disease. Journal of medical genetics. 1994;31(11): 872-874.

36. Brandt J, Bylsma FW, Gross R, Stine OC, Ranen N, Ross CA. Trinucleotide repeat length and clinical progression in Huntington's disease. Neurology. 1996;46(2): 527-531.

37. Tabrizi SJ, Scahill RI, Owen G, Durr A, Leavitt BR, Roos RA, et al. Predictors of phenotypic progression and disease onset in premanifest and early-stage Huntington's disease in the TRACK-HD study: analysis of 36-month observational data. Lancet Neurol. 2013;12: 637-649. doi:10.1016/S1474-4422(13)70088-7

38. Næs T, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. J Chemom. 2001;15(4): 413-26.

39. Vivot A, Glymour MM, Tzourio C, Amouyel P, Chêne G, Dufouil C. Association of Alzheimer's related genotypes with cognitive decline in multiple domains: results from the Three-City Dijon study. Molecular psychiatry. 2015;20(10): 1173-1178.

40. Mura T, Proust-Lima C, Jacqmin-Gadda H, Akbaraly TN, Touchon J, Dubois B, et al. Measuring cognitive change in subjects with prodromal Alzheimer's disease. Journal of Neurology, Neurosurgery & Psychiatry. 2014;85(4): 363-70.

41. Biglan KM, Shoulson I, Kieburtz K, Oakes D, Kayson E, Shinaman MA, et al. Clinical-Genetic Associations in the Prospective Huntington at Risk Observational Study (PHAROS): Implications for Clinical Trials. JAMA neurology. 2016;73(1): 102-110.

42. Epping EA, Kim JI, Craufurd D, Brashers-Krug TM, Anderson KE, McCusker E, et al. Longitudinal Psychiatric Symptoms in Prodromal Huntington's Disease: A Decade of Data. American Journal of Psychiatry. 2016;173(2): 184-92.

43. Vuono R, Winder-Rhodes S, De Silva R, Cisbani G, Drouin-Ouellet J, Spillantini MG et al. The role of tau in the pathological process and clinical expression of Huntington's disease. Brain. 2015;138(7): 1907-1918.

611 44. Williams-Gray CH, Hampshire A, Barker RA, Owen AM. Attentional control in
612 Parkinson's disease is dependent on COMT val158met genotype. Brain. 2008;131:
613 397-408. doi:10.1093/brain/awm313

614 45. Gennatas ED, Cholfin JA, Zhou J, Crawford RK, Sasaki DA, Karydas A, et al. COMT
615 Val158Met genotype influences neurodegeneration within dopamine-innervated brain
616 structures. Neurology. 2012;78: 1663-1669. doi:10.1212/WNL.0b013e3182574fa1

617 46. Frank MJ, Fossella JA. Neurogenetics and pharmacology of learning, motivation, and
618 cognition. Neuropsychopharmacology. Nature Publishing Group. 2011;36: 133-152.
619 doi:10.1038/npp.2010.96

620 47. Charvin D, Roze E, Perrin V, Deyts C, Betuing S, Pagès C, et al. Haloperidol protects
621 striatal neurons from dysfunction induced by mutated huntingtin in vivo. Neurobiol
622 Dis. 2008;29: 22-29. doi:10.1016/j.nbd.2007.07.028

623 48. Schwab LC, Garas SN, Drouin-Ouellet J, Mason SL, Stott SR, Barker RA. Dopamine
624 and Huntington's disease. Expert Rev Neurother. 2015;15: 445-458.
625 doi:10.1586/14737175.2015.1025383

626

Supporting information

S-Table 1: Demographic characteristics and performance of HD gene carriers including in the longitudinal analysis (N=350).

	Met / Met N=79	Met / Val N=175	Val / Val N=96	*p-value**
Age (yrs)	47.49 (12.8)	50.6 (11.6)	49.6 (10.4)	Ns
Sex (% men)	60.8	48.6	53.1	Ns
Age at onset (yrs)	41.6 (11.1)	45.2 (11.3)	43.6 (9.7)	Ns
Educational level (yrs in education)	12.2 (3.4)	11.1 (2.8)	12.5 (3.5)	0.0010
BMI	22.8 (4.0)	22.7 (3.5)	22.3 (3.6)	Ns
CAG repeats *mHtt*	45.4 (4.8)	44.3 (3.6)	44.5 (3.2)	Ns
CAG repeats *Htt*	18.0 (2.6)	18.9 (4.1)	18.9 (4.1)	Ns
UHDRS				
Motor	34.8 (17.3)	34.3 (20.7)	38.2 (21.7)	Ns
Behavior	19.3 (14.3)	17.8 (10.8)	16.9 (12.2)	Ns
FAS	29.7 (4.8)	29.9 (5.5)	31.0 (6.0)	Ns
IS	82.7 (13.5)	82.4 (15.8)	79.9 (16.6)	Ns
TFC	9.0 (3.2)	9.3 (3.2)	8.7 (3.5)	Ns
L Fluency 1’	20.2 (11.1)	19.2 (11.4)	17.2 (11.9)	Ns
L Fluency 2’	28.9 (17.4)	27.5 (17.7)	24.0 (18.2)	Ns
Stroop W	58.1 (21.2)	60.5 (22.1)	60.8 (26.8)	Ns
Stroop C	42.3 (16.4)	42.2 (15.5)	42.7 (18.7)	Ns
Stroop W/C	21.3 (10.0)	22.6 (11.7)	20.2 (13.4)	Ns
SDMT	20.7 (10.4)	22.2 (13.7)	21.2 (14.2)	Ns

HD: Huntington’s disease; BMI: body mass index; CAG repeats refers to the number of CAGs in the mutated (*mHtt*) and non-mutated (normal *Htt*) alleles of the Huntingtin gene; UHDRS: Unified Huntington’s Disease Rating Scale; TMS: Total Motor Score; FAS: Functional Assessment Scale; IS: Independence Scale; TFC: Total Functional Capacity; Letter fluency (L Fluency) was tested with PRV letters (French norms) at 1 minute (1’) and 2 minutes (2’); Stroop C: Color; W: Word; W/C: Word/Color (interference score); SDMT: Symbol Digit Modalities Test. Quantitative variables are presented as means, with the standard deviation in brackets, and qualitative variables are presented as frequency counts. Medication use is expressed as a percentage.*Non corrected *P*-values; Chi-squared test for qualitative variables and one-way ANOVA for quantitative data; Ns: not significant.

S-Table 2: Modelling results of the sensitivity analysis excluding outliers

	Motor	(N=312)	Behavior	(N=312)	Functional	(N=312)	Cognitive	(N=308)
	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)
Baseline:								
Met/Val vs Met/Met	0.24 (0.19)	0.1944 (ns)	-0.42 (0.38)	0.2722 (ns)	-0.01 (0.20)	0.9449 (ns)	-0.13 (0.17)	0.4385 (ns)
Val/Val vs Met/Met	-0.15 (0.21)	0.4612 (ns)	-0.17 (0.36)	0.6414 (ns)	-0.11 (0.21)	0.5926 (ns)	-0.25 (0.19)	0.1845 (ns)
Val/Val vs Met/Val	-0.40 (0.17)	0.0216* (ns)	0.25 (0.25)	0.3184 (ns)	-0.10 (0.16)	0.5565 (ns)	-0.12 (0.15)	0.4504 (ns)
Number of CAG repeats	0.02 (0.03)	0.5846 (ns)	0.07 (0.19)	0.7266 (ns)	0.09 (0.08)	0.2930 (ns)	0.04 (0.03)	0.1492 (ns)
Education level	0.04 (0.02)	0.0307* (ns)	0.05 (0.02)	0.0292* (ns)	0.06 (0.02)	0.0024** (0.0096**)	0.09 (0.02)	<0.0001*** (<0.0001***)
Slope:								
Met/Val vs Met/Met	-0.03 (0.01)	0.0421* (ns)	0.07 (0.03)	0.0269* (ns)	-0.002 (0.01)	0.8822 (ns)	0.02 (0.01)	0.0224* (ns)
Val/Val vs Met/Met	0.01 (0.02)	0.3420 (ns)	0.04 (0.03)	0.2003 (ns)	-0.003 (0.01)	0.8027 (ns)	0.02 (0.01)	0.0294* (ns)
Val/Val vs Met/Val	0.04 (0.01)	0.0009*** (0.0054**)	-0.03 (0.03)	0.2374 (ns)	-0.002 (0.01)	0.8889 (ns)	0.001 (0.01)	0.9346 (ns)
Number of CAG repeats	-0.02 (0.002)	<0.0001*** (<0.0001***)	-0.01 (0.02)	0.6992 (ns)	-0.02 (0.004)	<0.0001*** (<0.0001***)	-0.02 (0.002)	<0.0001*** (<0.0001***)

The motor domain was modeled including the performances at TMS; the behavioral domain was modeled including the performances at behavior task of the UHDRS; the functional domain was modeled including the performances at FAS and IS (TFC could not be included because there are not enough values for the model to converge); the cognitive domain was modeled including performances at letter fluency assessed at 1 and 2 minutes, SDMT and the three parts of the Stroop. N: Number of HD gene carriers who have contributed to the estimation (cognitive tasks were not available for all HD gene carriers); SE: Standard error of the estimate, *P*: *P*-values (*** P<0.001, ** P<0.01, *P<0.05). *Baseline* values correspond to the impact of covariates at estimated age at onset. *Slope* values correspond to the impact of covariates on the slope of the decline.

S-Table 3: Modelling results of linear mixed models for each task

	Motor		Behavior		Functional			
	TMS	(N=348)	Behavior	(N=348)	IS	(N=348)	FAS	(N=348)
	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)
Baseline:								
Met/Val vs Met/Met	-0.42 (3.27)	0.8976 (ns)	2.53 (1.98)	0.2034 (ns)	-2.57 (2.65)	0.3333 (ns)	1.64 (1.01)	0.1049 (ns)
Val/Val vs Met/Met	4.44 (3.67)	0.2275 (ns)	1.26 (2.20)	0.5662 (ns)	-3.41 (2.97)	0.2514 (ns)	1.09 (1.13)	0.3345 (ns)
Val/Val vs Met/Val	4.86 (3.05)	0.1122 (ns)	-1.26 (1.81)	0.4851 (ns)	-0.85 (2.46)	0.7306 (ns)	-0.55 (0.93)	0.5593 (ns)
Number of CAG repeats	-0.09 (0.33)	0.7801 (ns)	-0.43 (0.21)	0.0370* (ns)	0.81 (0.27)	0.0026** (0.0260*)	-0.37 (0.10)	0.0003** (0.0030**)
Education level	-0.51 (0.33)	0.1176 (ns)	-0.42 (0.16)	0.0075** (ns)	0.72 (0.25)	0.0039** (0.0390*)	-0.26 (0.10)	0.0073** (ns)
Slope:								
Met/Val vs Met/Met	0.17 (0.24)	0.4700 (ns)	-0.42 (0.19)	0.0247* (ns)	0.28 (0.21)	0.1978 (ns)	-0.16 (0.08)	0.0468* (ns)
Val/Val vs Met/Met	-0.27 (0.27)	0.3143 (ns)	-0.22 (0.21)	0.2824 (ns)	0.12 (0.24)	0.6300 (ns)	0.001 (0.09)	0.9881 (ns)
Val/Val vs Met/Val	-0.44 (0.22)	0.0476* (ns)	0.20 (0.17)	0.2455 (ns)	-0.16 (0.20)	0.4224 (ns)	0.16 (0.07)	0.0312* (ns)
Number of CAG repeats	0.18 (0.02)	<0.0001*** (<0.0001***)	0.04 (0.02)	0.0619 (ns)	-0.17 (0.02)	<0.0001*** (<0.0001***)	0.07 (0.01)	<0.0001*** (<0.0001***)

S-Table 3 continued

	Cognitive					
	Letter Fluency 1'	(N=338)	Letter Fluency 2'	(N=339)	SDMT	(N=321)
	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)
Baseline:						
Met/Val vs Met/Met	-3.49 (2.09)	0.0962 (ns)	-6.30 (2.98)	0.0355* (ns)	-0.30 (2.08)	0.8843 (ns)
Val/Val vs Met/Met	-5.68 (2.32)	0.0146* (ns)	-9.75 (3.31)	0.0034** (ns)	-0.14 (2.28)	0.9496 (ns)
Val/Val vs Met/Val	-2.19 (1.95)	0.2627 (ns)	-3.45 (2.78)	0.2159 (ns)	0.16 (1.91)	0.9339 (ns)
Number of CAG repeats	-0.13 (0.22)	0.5422 (ns)	-0.36 (0.31)	0.2514 (ns)	0.01 (0.22)	0.9569 (ns)
Education level	0.94 (0.20)	<0.0001* (<0.0001*)	1.68 (0.28)	<0.0001* (<0.0001*)	0.76 (0.20)	0.0002*** (0.0020**)
Slope:						
Met/Val vs Met/Met	0.46 (0.17)	0.0058** (ns)	0.83 (0.24)	0.0008*** (0.0120*)	0.18 (0.16)	0.2718 (ns)
Val/Val vs Met/Met	0.56 (0.18)	0.0021** (0.0315*)	0.87 (0.27)	0.0017** (0.0255*)	-0.15 (0.17)	0.3790 (ns)
Val/Val vs Met/Val	0.10 (0.16)	0.5089 (ns)	0.04 (0.24)	0.8633 (ns)	-0.33 (0.15)	0.0271* (ns)
Number of CAG repeats	-0.06 (0.02)	0.0013** (0.0130*)	-0.06 (0.03)	0.0166* (ns)	-0.05 (0.02)	0.0101* (ns)

S-Table 3 continued

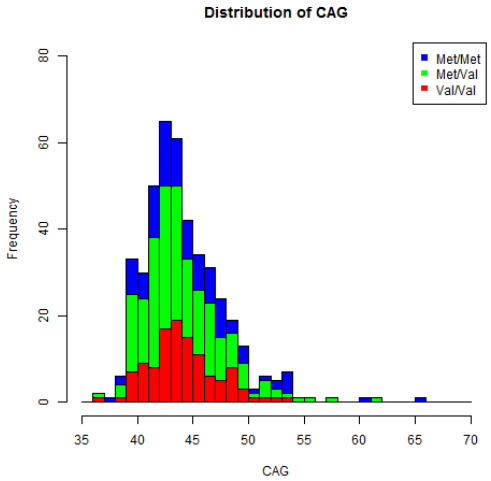
	Cognitive					
	Stroop C	(N=329)	Stroop W	(N=328)	Stroop C/W	(N=325)
	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)	Estimate (SE)	*P* (corrected *P*)
Baseline:						
Met/Val vs Met/Met	-3.02 (2.72)	0.2679 (ns)	-1.74 (3.68)	0.6365 (ns)	0.13 (1.87)	0.9449 (ns)
Val/Val vs Met/Met	-3.39 (3.05)	0.2667 (ns)	-3.82 (4.13)	0.3552 (ns)	-3.31 (2.10)	0.1162 (ns)
Val/Val vs Met/Val	-0.37 (2.52)	0.8830 (ns)	-2.08 (3.41)	0.5423 (ns)	-3.44 (1.74)	0.0486* (ns)
Number of CAG repeats	0.61 (0.27)	0.0250* (ns)	0.03 (0.37)	0.9285 (ns)	0.52 (0.19)	0.0056** (ns)
Education level	0.81 (0.27)	0.0033** (0.0330*)	1.20 (0.37)	0.0011** (0.0110*)	0.61 (0.19)	0.0012** (0.0120*)
Slope:						
Met/Val vs Met/Met	0.43 (0.20)	0.0316* (ns)	0.44 (0.28)	0.1105 (ns)	0.13 (0.14)	0.3663 (ns)
Val/Val vs Met/Met	0.22 (0.23)	0.3356 (ns)	0.48 (0.31)	0.1235 (ns)	0.39 (0.16)	0.0157* (ns)
Val/Val vs Met/Val	-0.21 (0.19)	0.2635 (ns)	0.04 (0.26)	0.8672 (ns)	0.26 (0.13)	0.0497* (ns)
Number of CAG repeats	-0.13 (0.02)	<0.0001*** (<0.0001***)	-0.12 (0.03)	<0.0001*** (<0.0001***)	-0.05 (0.01)	0.0001*** (0.0010**)

TMS: Total motor score, IS: Independence Scale, FAS: Functional Assessment Scale, SDMT: Symbol Digit Modalities Test, Stroop C: Stroop Color, Stroop W: Stroop Word, Stroop W/C: Stroop interference.

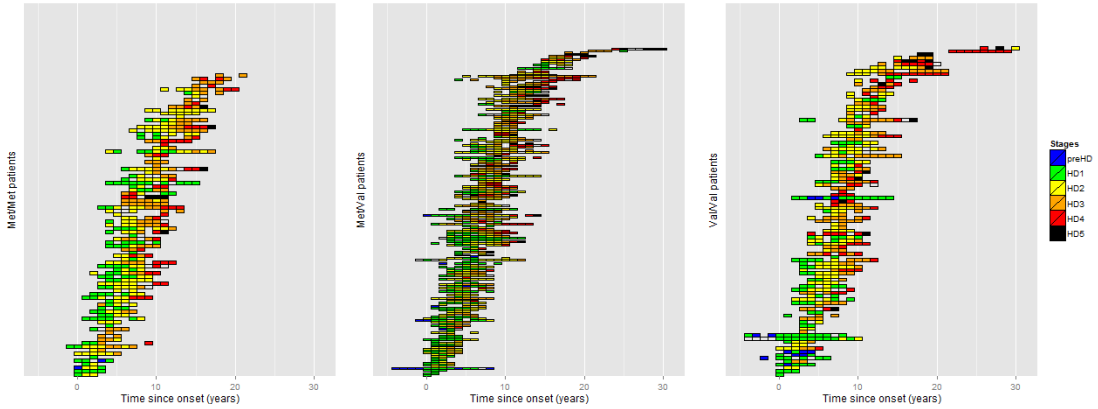
N: Number of HD gene carriers who have contributed to the estimation (cognitive tasks were not available for all HD gene carriers); SE: Standard error of the estimate, *P*: *P*-values (*** P<0.001, ** P<0.01, *P<0.05).

Baseline values correspond to the impact of covariates at estimated age at onset. *Slope* values correspond to the impact of covariates on the slope of the decline.

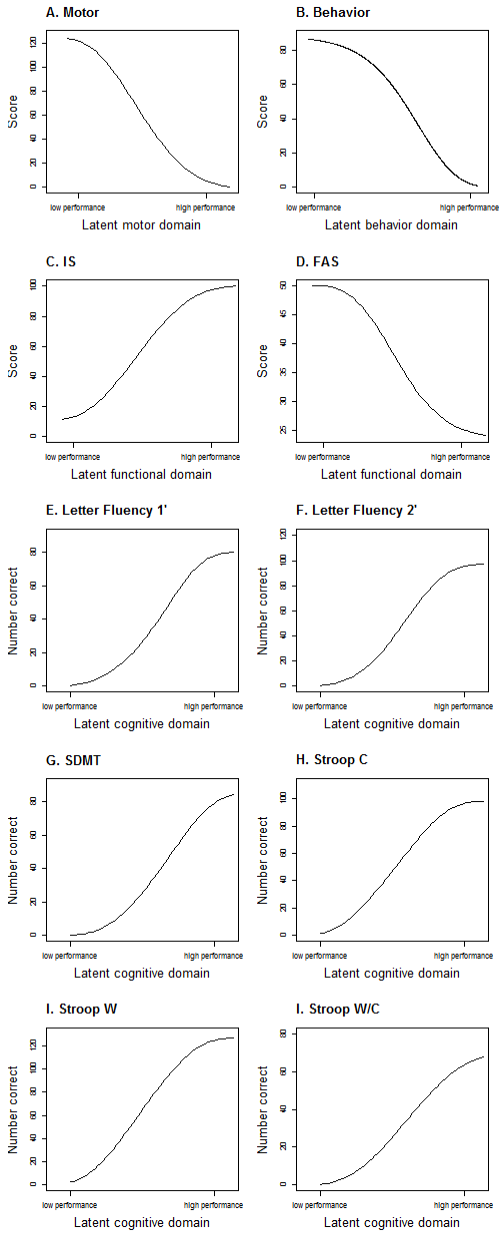
S1 Figure: Distribution of CAG repeats length in the database.



S2 Figure: Repartition of stages in time according to COMT polymorphisms.



S3 Figure: Link functions between performances at each task and latent processes modelling the domains.



We plotted the link function between each task and latent domains. UHDRS motor score **(A)**; UHDRS behavioral **(B)**, IS: Independence Score **(C)**; FAS: Functional Assessment Scale **(D)**, cognitive (letter fluency 1': at 1 minute **(E)**; letter fluency 2': at 2 minutes **(F)**; SDMT: symbol digit modalities test **(G)**; Stroop C: Stroop color **(H)**; Stroop W: Stroop word **(I)**; Stroop W/C: Stroop interference **(J)**.

5.2 Exemples d'intégration des marqueurs pronostiques dans les essais cliniques

Outre dans les soins courants, les marqueurs pronostiques peuvent être pris en compte dans les essais cliniques afin de les améliorer. Ils peuvent par exemple être utilisés comme critère d'inclusion ou de stratification ou encore comme variable d'ajustement.

• Les marqueurs pronostiques comme critère d'inclusion

Utiliser un marqueur pronostique comme critère d'inclusion permet d'inclure un groupe plus homogène de patients. Dans le cadre d'un essai clinique de petit effectif, réduire la variabilité inter-patients est un moyen d'augmenter la puissance de l'étude. Le polymorphisme COMT pourrait être utilisé comme critère d'inclusion dans un essai clinique portant sur un critère cognitif longitudinal. Les résultats ne pourront pas être généralisés aux patients ne présentant pas le marqueur. Par exemple, pour le polymorphisme COMT, réaliser un essai clinique n'incluant que les patients homozygotes Val/Val ne permet pas de tirer de conclusion quant aux patients Met/Met ou Met/Val.

• Les marqueurs pronostiques comme critère de stratification

Utiliser un marqueur pronostique comme critère de stratification permet d'équilibrer les bras de traitement sur le marqueur. Bien que la randomisation permet d'équilibrer les groupes, dans le cadre d'un essai clinique de petit effectif, il est possible que des groupes déséquilibrés apparaissent. Dans ce cas, les bras de traitement pourraient différer sur la sévérité de la maladie et/ou son évolution et fausser l'interprétation de l'essai clinique.

• Les marqueurs pronostiques comme variable d'ajustement

Enfin, lorsque le marqueur pronostique n'a pas été pris en compte dans la conception de l'essai clinique, il peut être utilisé comme variable d'ajustement. Si cela n'a pas été prévu dans l'analyse, il peut s'agir d'une analyse de sensibilité qui viendra compléter l'analyse principale. Les conclusions de cette analyse permettront d'estimer la part expliquée par le traitement et celle expliquée par le marqueur pronostique.

Chapitre 6

Intégration des marqueurs prédictifs dans les essais cliniques

Pour qu'un marqueur prédictif puisse être utilisé en soins courants pour déterminer le traitement le mieux adapté au patient, son utilité clinique doit être démontrée grâce à une étude prospective. Il faut prouver que le traitement correspondant est le plus efficace pour le groupe $M+$ et que ce traitement n'a pas d'intérêt pour le groupe $M-$. Les essais cliniques intégrant un marqueur prédictif dans le plan expérimental permettent de répondre à ces questions. Nous avons réalisé une revue de la littérature afin de définir les conditions d'utilisation de chaque plan expérimental. Nous avons aussi réalisé une étude de simulation pour définir leur puissance et leurs limites.

6.1 Les plans expérimentaux d'essai clinique basés sur un marqueur prédictif (Etat de l'art)

Plusieurs plans expérimentaux basés sur les marqueurs prédictifs ont émergé ces vingt dernières années, en particulier dans le cadre de l'oncologie [108, 109, 110, 111, 112]. Ces plans permettent de déterminer si le traitement est efficace dans la sous-population $M+$, et/ou si le marqueur est réellement prédictif de l'efficacité d'un traitement (plans indirects) et/ou si la stratégie consistant à donner le traitement seulement aux patients $M+$ est la meilleure (plans directs).

Suite à une étude bibliographique, nous avons sélectionné six plans expérimentaux se basant sur un marqueur prédictif, et construits à partir d'un plan parallèle de comparaison de deux traitements : le traitement expérimental dont l'effet est dépendant du marqueur M et le traitement standard (Figure 26).

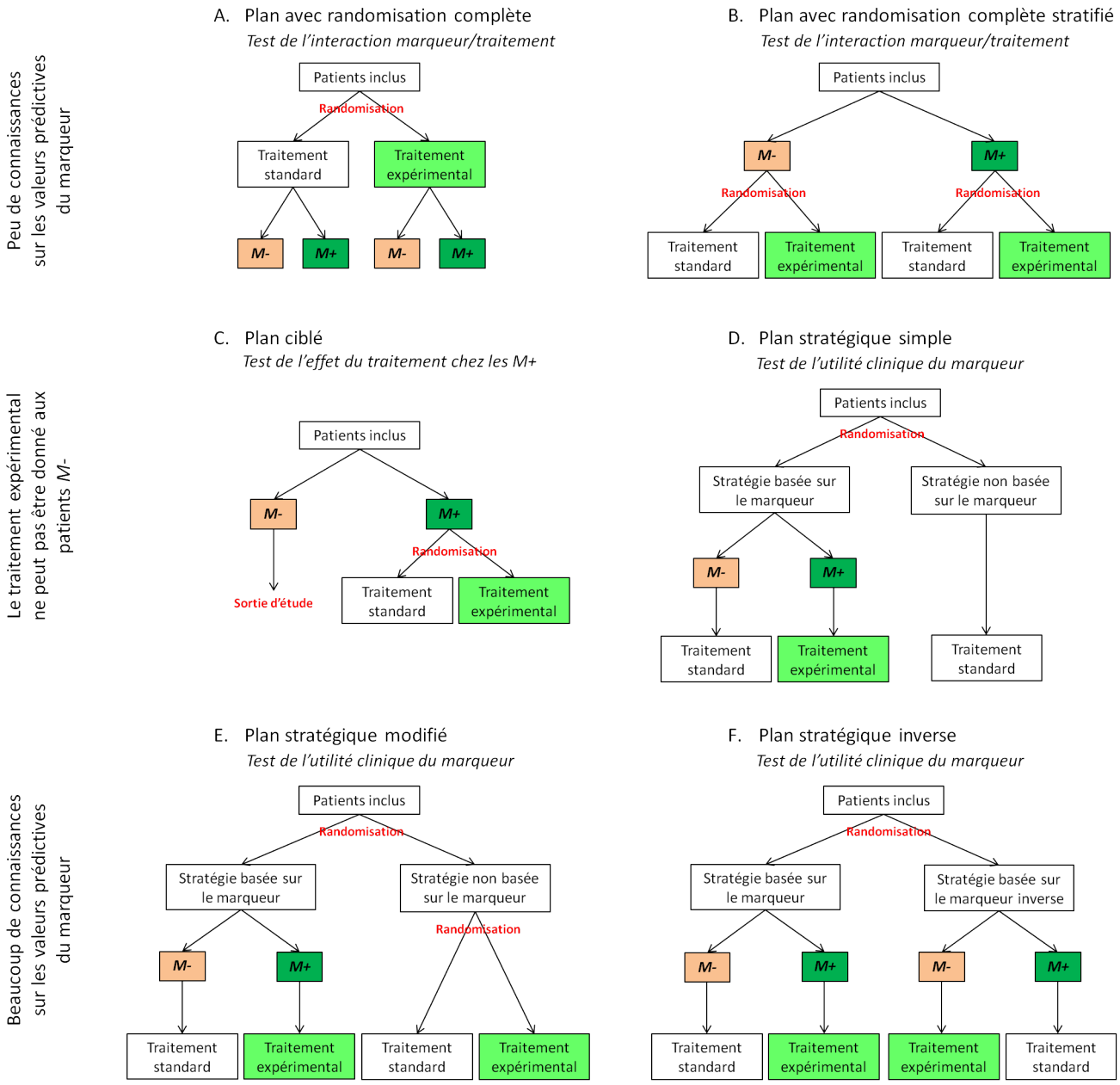


FIGURE 26 – Représentation schématique des plans expérimentaux basés sur un marqueur prédictif

Les plans A et B sont identiques à l'exception du fait que le plan A randomise tous les patients en une seule fois, et le plan B randomise chaque sous-groupe $M+$ et $M-$. Le plan C ne randomise que les patients $M+$. Cela correspond à un critère d'inclusion supplémentaire. Les plans D, E et F testent l'utilité de la mise en place d'une médecine stratifiée basée sur le marqueur M pour le traitement expérimental.

• Les plans avec randomisation complète permettent de tester l'interaction entre marqueur et traitement

Le plan avec **randomisation complète** (« *Randomize-all design* ») correspond au plan parallèle standard pour lequel on ajoute une phase de détermination de la valeur du marqueur M (Figure 26.A). Il est utilisé essentiellement lors d'études prospectives visant à étudier l'effet du traitement dans la population totale. Le marqueur fait alors partie des variables étudiées et sera utilisé dans des études rétrospectives ou dans des analyses en sous-groupes. Si ces analyses en sous-groupes n'ont pas été prévues dans le protocole, leur étude peut pâtir d'un manque de puissance. De plus, déterminer le statut du marqueur à la fin ou au cours de l'essai ne garantit pas l'exhaustivité des données. Le plan avec **randomisation complète stratifiée ou plan d'interaction** (« *marker-stratified design* » ou « *marker-interaction design* ») pallie ce problème en déterminant le statut du marqueur avant les randomisations indépendantes des groupes $M+$ et $M-$ (Figure 26.B). Evidemment, cela n'est possible qu'à condition que le processus de détermination du statut du marqueur ne soit pas long devant l'urgence de la mise en place du traitement. Lorsqu'il s'agit d'études prospectives, différentes analyses statistiques peuvent être planifiées : (i) test de l'efficacité du traitement expérimental dans chaque sous-groupe $M+$ et $M-$, (ii) test de l'interaction entre le marqueur et le traitement, puis test de l'effet du traitement dans le sous-groupe $M+$ si interaction significative, (iii) test de l'effet du traitement chez tous les patients, puis test de l'effet du traitement dans le sous-groupe $M+$ si non significatif ; l'analyse (ii) étant celle avec la plus grande puissance [113]. Cependant le choix de l'analyse reste conditionné par les connaissances et intuitions sur le caractère prédictif du marqueur [114]. Ces plans permettent de conclure à la valeur prédictive du marqueur M sur l'effet du traitement expérimental.

• Le plan ciblé n'inclut que les patients $M+$

Le plan **ciblé ou enrichi** (« *Targeted design* » ou « *Enrichment design* ») ne randomise que le sous-groupe $M+$ des patients inclus (Figure 26.C). En présence d'un marqueur prédictif, ce plan nécessite de randomiser moins de patients par rapport à un plan parallèle classique quand le test diagnostique associé au marqueur a de fortes spécificité et sensibilité et que le traitement expérimental n'a aucun effet chez les patients $M-$. Cependant, le nombre de patients à inclure augmente lorsque lorsque la prévalence de $M+$ est faible [115, 116]. Ce plan ne permet pas, ni d'évaluer la valeur prédictive du marqueur M, ni l'effet du traitement chez les patients $M-$. L'utilisation de ce plan expérimental en phase III nécessite qu'il y ait une forte évidence que le traitement expérimental ne peut pas avoir d'effet bénéfique pour les patients $M-$ [114].

• Les plans stratégiques évaluent directement la stratégie de médecine stratifiée

Ces plans expérimentaux cherchent, indirectement, à évaluer la stratégie de médecine stratifiée consistant à donner le traitement en fonction du marqueur. On peut les opposer à des plans dits « directs » comparant deux stratégies de choix du traitement [108]. La stratégie mise en avant consiste à donner le traitement expérimental au groupe $M+$ et le traitement standard au groupe $M-$. Elle est comparée à une autre stratégie qui ne s'appuie pas sur le marqueur pour donner le traitement. Là encore la connaissance du marqueur doit se faire avant la randomisation du patient afin de s'assurer de la connaissance du statut du patient. Dans le plan **stratégique simple** (« *Marker-based strategy design* »), la stratégie « contrôle » traite tous les patients avec le traitement standard (Figure 26.D). Le plan **stratégique modifié** (« *Modified marker-based strategy design* ») reprend le principe du plan stratégique simple en permettant aux patients $M-$ d'avoir accès au traitement expérimental. La stratégie « contrôle » randomise les patients indépendamment du marqueur dans le bras de traitement standard ou expérimental (Figure 26.E). Ce plan ne peut être utilisé qu'à condition que le traitement expérimental ne soit pas supposé délétère pour les patients $M-$.

Actuellement, il n'y a pas un plan qui soit plus efficient (d'un point de vue puissance et nombre de sujets nécessaires) que les autres dans toutes les situations données [117]. Eng a récemment proposé le plan **stratégique inverse** (« *Reverse marker-based strategy design* ») qui compare la stratégie basée sur le marqueur (identique aux plans précédents) à une autre stratégie aussi basée sur la marqueur mais totalement opposée à la première [111]. Ainsi la seconde stratégie, dite « contrôle », attribue le traitement expérimental aux patients $M-$ et le traitement standard aux patients $M+$ (Figure 26.F). Ce plan permet de tester la valeur prédictive du marqueur tout en permettant d'inclure moins de sujets que les autres plans stratégiques. Il oppose au maximum les stratégies envisagées, avec l'absence totale de recouvrement dans les deux bras de randomisation ce qui lui permet de mettre plus facilement en évidence une différence entre les deux stratégies.

Les trois plans stratégiques (simple, modifié et inverse) peuvent inciter à donner de fausses conclusions sur la valeur prédictive du marqueur lorsque la stratégie basée sur le marqueur se montre plus efficace que la stratégie contrôle. En effet, la prévalence du marqueur, son caractère éventuellement pronostique et l'efficacité du traitement chez les patients $M-$ peuvent aussi avoir un impact sur les résultats de l'essai clinique. Nous nous sommes intéressés à ces problématiques en réalisant une étude de simulation, présentée dans la section 6.2 suivante.

6.2 Impact des valeurs prédictives et pronostiques du marqueur sur les plans expérimentaux stratégiques : une étude de simulation

6.2.1 Objectif et notations

Nous nous sommes intéressés aux trois plans expérimentaux comparant deux stratégies d'attribution d'un traitement, en randomisant les patients dans deux bras de stratégie. Nous nommerons « bras stratégique », le bras de randomisation associé à la stratégie basée sur le marqueur prédictif où les patients $M+$ sont traités avec le traitement expérimental et les patients $M-$ avec le traitement standard. L'autre bras sera nommé « bras contrôle ». Les patients du bras contrôle sont soit tous traités avec le traitement standard (plan stratégique simple), soit randomisés entre le traitement expérimental et le traitement standard (plan stratégique modifié), soit reçoivent le traitement en fonction du marqueur, avec le choix inverse du bras stratégique (plan stratégique inverse). Nous avons comparé ces trois plans expérimentaux sur :

• l'évolution du **nombre de sujets nécessaires** à inclure selon la taille de l'effet et de la prévalence du marqueur.

Puis, nous nous sommes intéressés à l'impact de l'utilisation de ces plans expérimentaux en essai clinique sur la mise en place d'une stratégie appropriée lorsque la connaissance de l'interaction marqueur/traitement n'était pas suffisamment grande :

• le risque de conduire une médecine stratifiée pour le traitement expérimental alors qu'il serait bénéfique pour toute la population, c'est-à-dire la probabilité de montrer une différence entre les deux stratégies lorsque **le marqueur n'est pas prédictif** de l'efficacité du traitement expérimental ;

• le risque de ne pas conduire une médecine stratifiée pour le traitement expérimental alors qu'il n'est bénéfique qu'aux patients $M+$ suite à l'impact de **la valeur pronostique du marqueur** sur la mise en évidence d'une différence entre les deux stratégies.

Les valeurs prédictive et pronostique du marqueur sont deux points essentiels de la mise en œuvre d'un plan expérimental stratégique. En effet, ces deux paramètres influencent la puissance du test statistique correspondant à la comparaison des deux bras de stratégie. Une mauvaise connaissance de ceux-ci peut induire la mise en place de fausses stratégies de médecine stratifiée ou au contraire aboutir à une thérapie homogène pour tous les patients avec les effets de toxicité que cela comporte.

Dans le cadre des données longitudinales, avec un traitement tel que les greffes, le critère de jugement peut être binaire (« Le score du patient a-t-il baissé de $x\,\%$ en 6 mois ? oui/non ») ou continu (score du patient en fin d'étude, la randomisation garantissant la comparabilité des patients en début d'étude ; différentiel de score entre le début et la fin de l'étude). Nous présentons les résultats dans le cas d'un critère binaire.

Notations

Soit π la prévalence du marqueur $M+$. La probabilité p de réaliser l'événement (par exemple « guérison ») , conditionnellement au traitement T et au marqueur M peut s'écrire :

$$p = \theta_0 + \theta_{M+} \times \mathbb{1}_{(M=M+)} + \theta_{Exp} \times \mathbb{1}_{(T=Exp)} + \theta_{Std} \times \mathbb{1}_{(T=Std)} + \theta_{Exp+} \times \mathbb{1}_{(T=Exp,M=M+)} + \theta_{Std+} \times \mathbb{1}_{(T=Std,M=M+)} \quad (6.1)$$

où θ_0 correspond à la probabilité de faire l'événement pour un patient $M-$ ne recevant aucun traitement, θ_{M+} correspond à l'effet additionnel du marqueur $M+$ indépendamment du traitement (effet pronostique), θ_{Exp} et θ_{Sdt} correspondent aux effet additionnels du traitement expérimental et du traitement standard indépendamment du marqueur et θ_{Exp+} et θ_{Sdt+} correspondent aux effets d'interaction entre le traitement expérimental ou le traitement standard et le marqueur $M+$ (effets prédictifs).

De plus, nous utiliserons les notations moyennes suivantes :

- p_S : probabilité de faire l'événement pour un patient inclus dans le bras stratégique
- p_C : probabilité de faire l'événement pour un patient inclus dans le bras contrôle
- $p_{Exp+} = \theta_0 + \theta_{M+} + \theta_{Exp} + \theta_{Exp+}$: probabilité de faire l'événement pour un patient $M+$ recevant le traitement expérimental
- $p_{Exp-} = \theta_0 + \theta_{Exp}$: probabilité de faire l'événement pour un patient $M-$ recevant le traitement expérimental
- $p_{Std+} = \theta_0 + \theta_{M+} + \theta_{Std} + \theta_{Std+}$: probabilité de faire l'événement pour un patient $M+$ recevant le traitement standard
- $p_{Std-} = \theta_0 + \theta_{Std}$: probabilité de faire l'événement pour un patient $M-$ recevant le traitement standard.

La table 7 résume les valeurs prises par p_S et p_C en fonction de p_{Exp+}, p_{Exp-}, p_{Std+} et p_{Std-} pour chaque plan expérimental stratégique.

TABLE 7 – Probabilités théoriques p_S et p_C pour chaque plan expérimental stratégique

Plan	p_S	p_C
Stratégique simple	$\pi p_{Exp+} + (1-\pi)p_{Std-}$	$\pi p_{Std+} + (1-\pi)p_{Std-}$
Stratégique modifié	$\pi p_{Exp+} + (1-\pi)p_{Std-}$	$\frac{\pi}{2}(p_{Exp+} + p_{Std+}) + \frac{1-\pi}{2}(p_{Exp-} + p_{Sdt-})$
Stratégique inverse	$\pi p_{Exp+} + (1-\pi)p_{Std-}$	$\pi p_{Std+} + (1-\pi)p_{Exp-}$

Remarque : la stratégie basée sur le marqueur est la même dans chacun des plans expérimentaux.

Puissance du test

Les deux bras de stratégie sont comparés grâce au test bilatéral suivant :

$$H_0 : p_S = p_C \quad \text{versus} \quad H_1 : p_S \neq p_C \tag{6.2}$$

Soit n le nombre de patients dans chaque bras (hypothèse d'un ratio 1 : 1), α le risque de première espèce et β le risque de deuxième espèce (puissance : $1-\beta$) associés au test (6.2). Ces trois paramètres sont reliés par l'équation (6.3).

$$n = \frac{(Z_{\alpha/2} + Z_{1-\beta})^2 \ [p_S(1-p_S) + p_C(1-p_C)]}{(p_S - p_C)^2} \tag{6.3}$$

où Z_u correspond au $u^{\text{ième}}$ quantile de la loi normale centrée réduite. La puissance associée au test (6.2) est alors donnée par :

$$1 - \beta = \mathbb{P}\left(u \ \leq \ \frac{\sqrt{n/2}\,(p_S - p_C)}{\sqrt{p_S(1-p_S) + p_C(1-p_C)}} - Z_{\alpha/2} \right) , \qquad u \sim \mathcal{N}(0,1) \tag{6.4}$$

6.2.2 Nombre de sujets nécessaires

Le nombre de sujets à inclure dans l'étude dépend des paramètres θ_0, θ_{Exp}, θ_{Sdt}, θ_{Exp+}, θ_{Std+} et de la prévalence du marqueur π. Nous avons étudié l'impact de ces paramètres en faisant varier π dans les quatre cas suivants :

A. Le traitement expérimental a un effet bénéfique chez les patients $M+$ et aucun effet chez les patients $M-$. Le traitement standard n'a aucun effet quel que soit le statut du marqueur.

B. Le traitement expérimental a un effet bénéfique chez tous les patients, avec un meilleur effet chez les patients $M+$. Le traitement standard a un effet quel que soit le statut du marqueur.

C. Le traitement expérimental a un effet bénéfique chez les patients $M+$ et aucun effet chez les patients $M-$. Le traitement standard a un effet bénéfique chez les patients $M-$ et aucun effet sur les patients $M+$.

D. Le traitement expérimental a un effet bénéfique chez tous les patients, avec un meilleur effet chez les patients $M+$. Le traitement standard a un effet bénéfique chez les patients $M-$ et aucun effet sur les patients $M+$.

Le cas A correspond aux cas où aucun traitement n'existe et que le traitement expérimental est comparé à un placébo. Le cas B correspond au cas où un traitement existe, mais qu'un nouveau traitement peut être plus efficace que celui-ci pour la sous-population $M+$. Les cas C et D correspondent à des cas où un traitement existe et qu'il n'est bénéfique que dans la sous-population de patients au statut $M-$.

Le plan stratégique simple nécessite d'inclure moins de patients lorsque la prévalence du marqueur est élevée (Figure 27). En effet, plus la prévalence du marquer $M+$ est grande, et plus les deux bras de stratégie vont être différents. Pour les plans stratégiques modifié et inverse, le nombre de sujets nécessaires à inclure tend vers $+\infty$ pour π valant $r = \frac{p_{Exp-} - p_{Sdt-}}{p_{Exp+} - p_{Sdt+} + p_{Exp-} - p_{Sdt-}} = \frac{\Delta_-}{\Delta_+ + \Delta_-}$, où Δ_- (respectivement Δ_+) correspond à la différence d'effet du traitement dans le groupe $M-$ (respectivement $M+$). Dans les cas A et B, $r = 0$, tandis que dans les cas C et D, $r < 0$. Le cas $0 < r < 1$ correspond au cas où le traitement expérimental est meilleur que le traitement standard, indépendamment du marqueur, ce qui n'est pas un scénario plausible à la mise en place d'un plan expérimental stratégique. Le plan stratégique inverse nécessite toujours d'inclure moins de patients que le plan stratégique modifié. Pour $r < 0$, le plan stratégique inverse est toujours le moins coûteux en nombre de sujets nécessaires à inclure. Pour $r = 0$, le plan stratégique inverse nécessite d'inclure le même nombre de sujets que le plan stratégique simple.

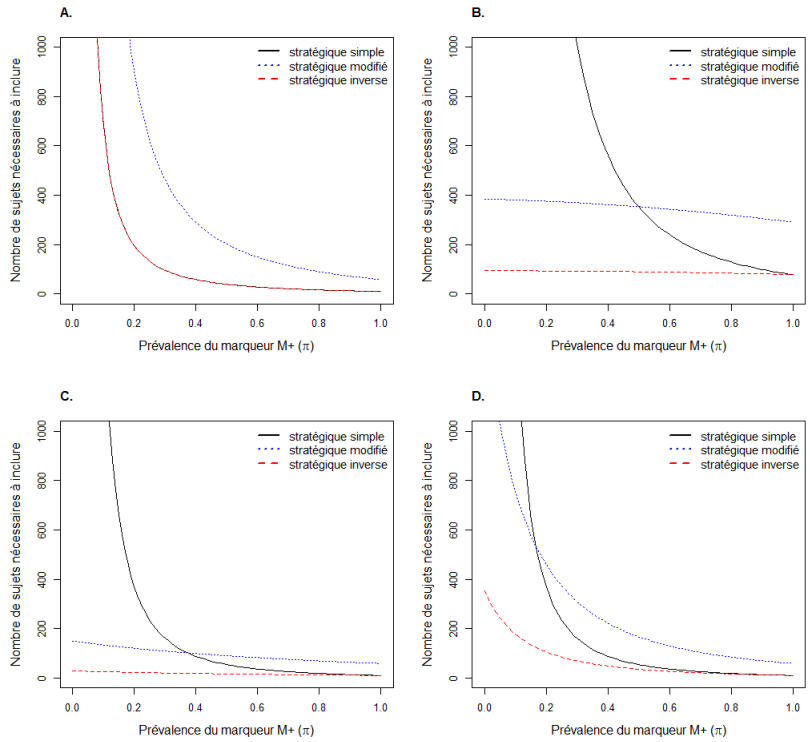


FIGURE 27 – Impact de la prévalence du marqueur $M+$ sur le nombre de sujets nécessaires à inclure dans l'étude

Les graphes ont été obtenus avec les paramètres suivants : $\theta_0 = 0{,}1$ et $\theta_{M+} = 0$. (A) $\theta_{Exp} = 0$, $\theta_{Exp+} = 0{,}5$, $\theta_{Sdt} = 0$ et $\theta_{Sdt+} = 0$. (B) $\theta_{Exp} = 0{,}3$, $\theta_{Exp+} = 0{,}4$, $\theta_{Sdt} = 0{,}5$ et $\theta_{Sdt+} = 0$. (C) $\theta_{Exp} = 0$, $\theta_{Exp+} = 0{,}5$, $\theta_{Sdt} = 0{,}3$ et $\theta_{Sdt+} = -0{,}3$. (D) $\theta_{Exp} = 0{,}2$, $\theta_{Exp+} = 0{,}3$, $\theta_{Sdt} = 0{,}3$ et $\theta_{Sdt+} = -0{,}3$.

6.2.3 Conséquences d'un marqueur prédictif et pronostique sur la puissance de l'étude

Un marqueur est à la fois prédictif et pronostique s'il influe sur l'efficacité du traitement et sur l'évolution de la maladie (Figure 23). La valeur pronostique du marqueur a un impact sur la puissance du test (6.2), et donc sur le nombre de sujets nécessaires. En effet, bien que θ_{M+} n'intervient pas dans le calcul de la différence d'effet du traitement entre les deux stratégies, il intervient dans le calcul de sa variance et la puissance (6.4) peut s'écrire, en fonction de la valeur pronostique du marqueur θ_{M+}, comme suit :

$$1 - \beta = \mathbb{P}\left(u \leq \frac{\sqrt{n}\gamma}{\sqrt{a + b\theta_{M+} + c\theta_{M+}^2}} - Z_{\alpha/2} \right), \qquad u \sim \mathcal{N}(0,1) \tag{6.5}$$

avec n le nombre de patients dans chaque bras de stratégie, et $(\gamma, a, b, c) \in \mathbb{R}^4$. L'équation (6.5) montre que la puissance est modifiée par la valeur pronostique du marqueur de façon non monotone. Nous avons évalué la puissance en fonction de la valeur pronostique pour différentes valeurs de n, de π et de θ_0. Nous nous sommes placés dans le cas où le

traitement expérimental a un effet bénéfique chez les patients M+ et aucun effet chez les patients M- tandis que le traitement standard n'a aucun effet quel que soit le statut du marqueur ($\theta_{Exp} = \theta_{Std} = \theta_{Std+} = 0$), correspondant au scénario A de la section 6.2.2. Dans ce cas, les valeurs des paramètres γ, a, b et c sont communes aux plans expérimentaux stratégiques simple et inverse (voir Annexe C pour les détails du calcul). Les résultats obtenus sont présentés sur les figures 28 et 29. La puissance est minimale pour θ_{M+} minimisant $\frac{\sqrt{n}\gamma}{\sqrt{a+b\theta_{M+}+c\theta_{M+}^2}}$, soit $\theta_{M+} = \frac{-b}{2c}$.

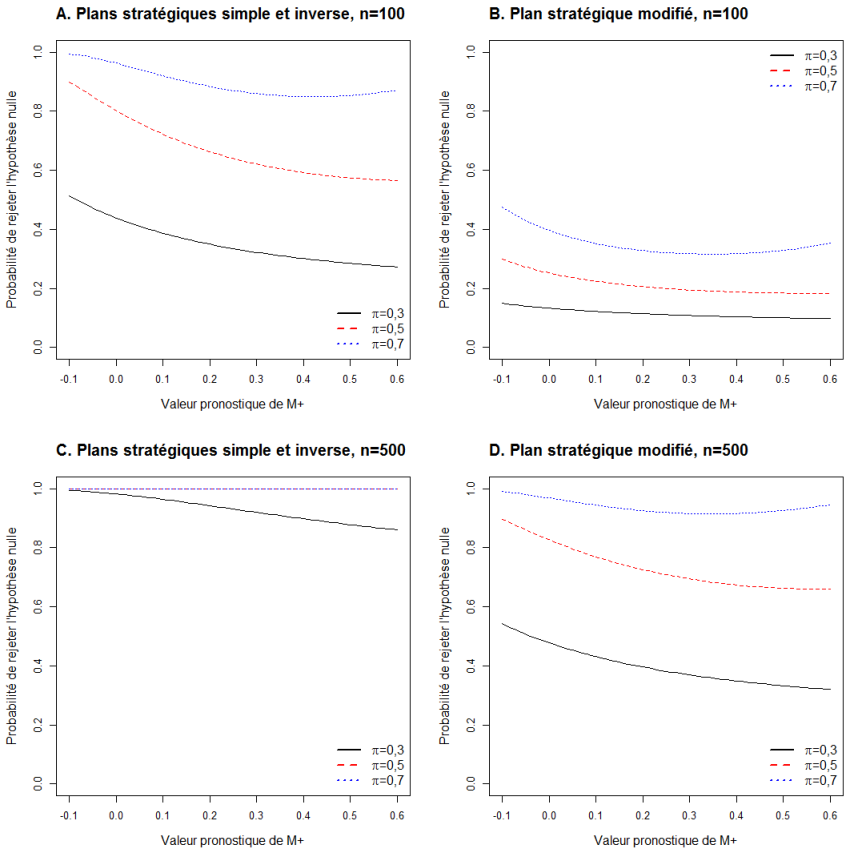


FIGURE 28 – Impact de la valeur pronostique du marqueur sur la puissance de l'étude pour différentes prévalences (π) du statut $M+$

Les graphes ont été obtenus avec les paramètres suivants : $\theta_0 = 0{,}1$, $\theta_{Epx} = 0$ et $\theta_{Exp+} = 0{,}3$.

Nos résultats montrent que l'impact de la valeur pronostique de $M+$ est modifiée en fonction de n, θ_0 et π. Nous aurions des résultats similaires avec le paramètre θ_{Exp+}. Ces résultats, obtenus dans un scénario très simple, se généralisent à d'autres scénarios, et ont pour but principal de montrer que la valeur prédictive du marqueur ne doit être sous-estimée, en particulier dans le cas de petits effectifs. En tenir compte dans le calcul du nombre de sujets nécessaires permet de garantir une puissance suffisante.

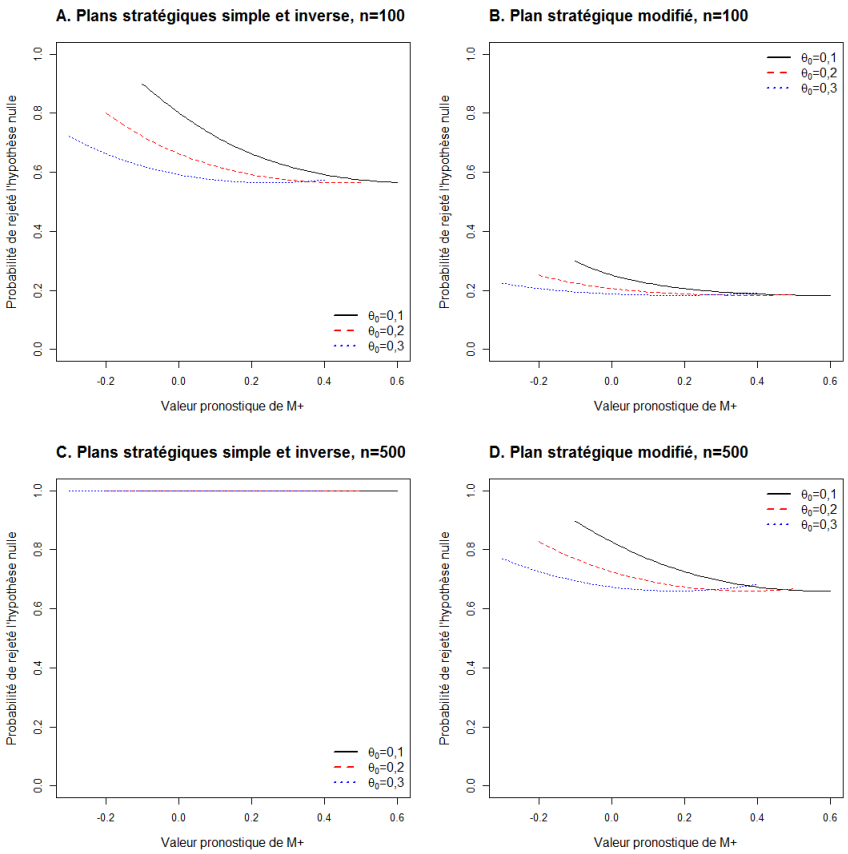


FIGURE 29 – Impact de la valeur pronostique du marqueur sur la puissance de l'étude pour différents effets de base (θ_0)

Les graphes ont été obtenus avec les paramètres suivants : $\pi = 0{,}5$, $\theta_{Epx} = 0$ et $\theta_{Exp+} = 0{,}3$.

6.2.4 Conséquences de l'utilisation de ces plans expérimentaux lorsque le traitement expérimental est meilleur que le traitement standard, indépendamment du marqueur

Lorsque le traitement expérimental est plus efficace que le traitement standard, indépendamment du marqueur, c'est toute la population qui doit en bénéficier et la mise en place d'un plan expérimental stratégique n'est pas considérée. Cependant, nous nous sommes demandés quelles seraient les conclusions suite à l'utilisation de ces plans dans ce cas, que l'hypothèse d'un marqueur prédictif de l'efficacité du traitement expérimental soit vérifiée ou non. Pour cela, nous avons calculé la probabilité de rejeter l'hypothèse H_0 associée au test (6.2) en fonction de la prévalence du marqueur $M+$ (π), de l'effet du traitement expérimental (θ_{exp}) et de la valeur prédictif du marqueur pour le traitement expérimental (θ_{exp+}). Les résultats obtenus sont présentés sur les figures 30 et 31.

Si le marqueur M n'est pas prédictif de l'effet du traitement

Lorsque le marqueur M n'est pas prédictif de l'effet du traitement expérimental ($\theta_{Exp+} = 0$), l'effet du traitement expérimental est le même pour tous les patients ($p_{Exp+} = p_{Exp-}$).

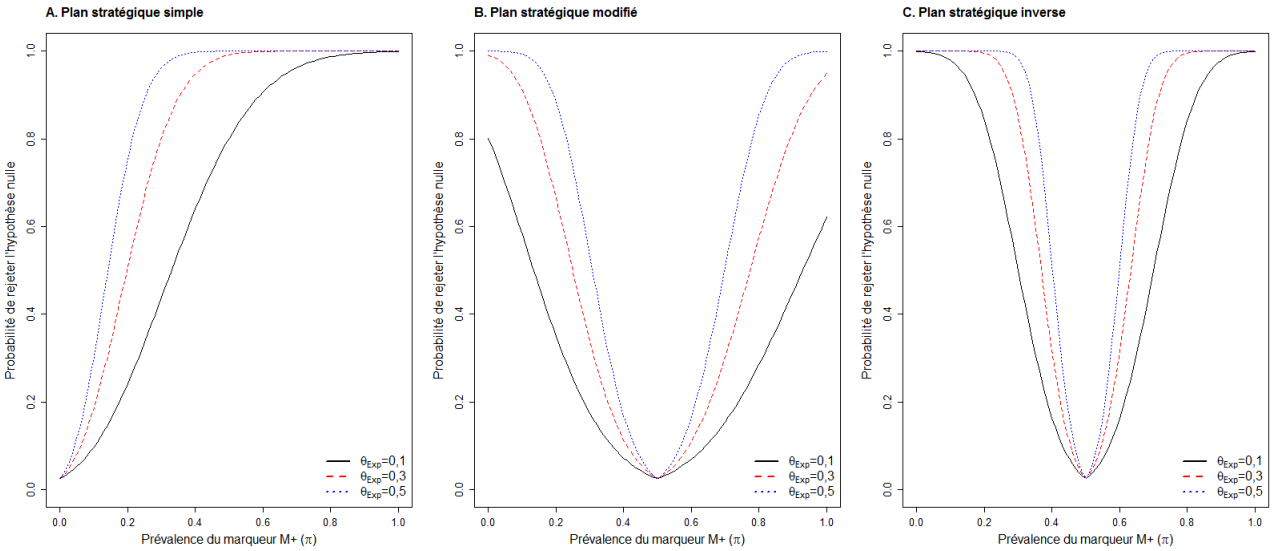


FIGURE 30 – Impact de la prévalence du marqueur $M+$ sur la probabilité de montrer une différence entre les deux stratégies lorsque le marqueur n'est pas prédictif de l'effet du traitement expérimental, pour différentes valeurs d'effet du traitement

Les graphes ont été obtenus avec les paramètres suivants : $\theta_0 = 0{,}1$, $\theta_{M+} = 0$, $\theta_{Std} = 0$, $\theta_{Std+} = 0$ et $\theta_{Exp+} = 0$ pour 100 patients dans chaque groupe de stratégie.

Pour chaque plan expérimental stratégique, la probabilité de rejeter l'hypothèse H_0 augmente lorsque l'effet du traitement augmente. Avec le plan stratégique simple, lorsque la prévalence π du marqueur $M+$ augmente, la probabilité de rejeter H_0 augmente et est toujours en faveur du bras stratégique. Avec les plans stratégiques modifié et inverse, H_0 est rejetée dans 5% des cas lorsque $\pi = 0{,}5$, et la probabilité de rejeter H_0 augmente, en faveur du bras stratégique lorsque π augmente, et en faveur du bras contrôle lorsque π diminue. Pour le plan stratégique inverse, la probabilité de conclure à une différence entre les deux stratégies est complètement symétrique par rapport à la prévalence du marqueur $M+$. Cela est dû à la conception symétrique de ce plan expérimental impliquant que la différence entre les deux stratégies et la variance associée sont les mêmes pour π et $1 - \pi$. Pour le plan stratégique modifié, il n'y a pas de symétrie parfaite car si la différence entre les deux stratégies est la même pour π et $1 - \pi$, la variance associée diffère de $(\pi - \frac{1}{2})(\theta_{Exp} - \theta_{Exp}^2) + (1 - 2\pi)\theta_0\theta_{Exp}$.

Si le marqueur M est prédictif de l'effet du traitement

Lorsque le marqueur M est prédictif de l'effet du traitement expérimental ($\theta_{Exp+} > 0$), l'effet du traitement expérimental est le même pour tous les patients ($p_{Exp+} > p_{Exp-}$).

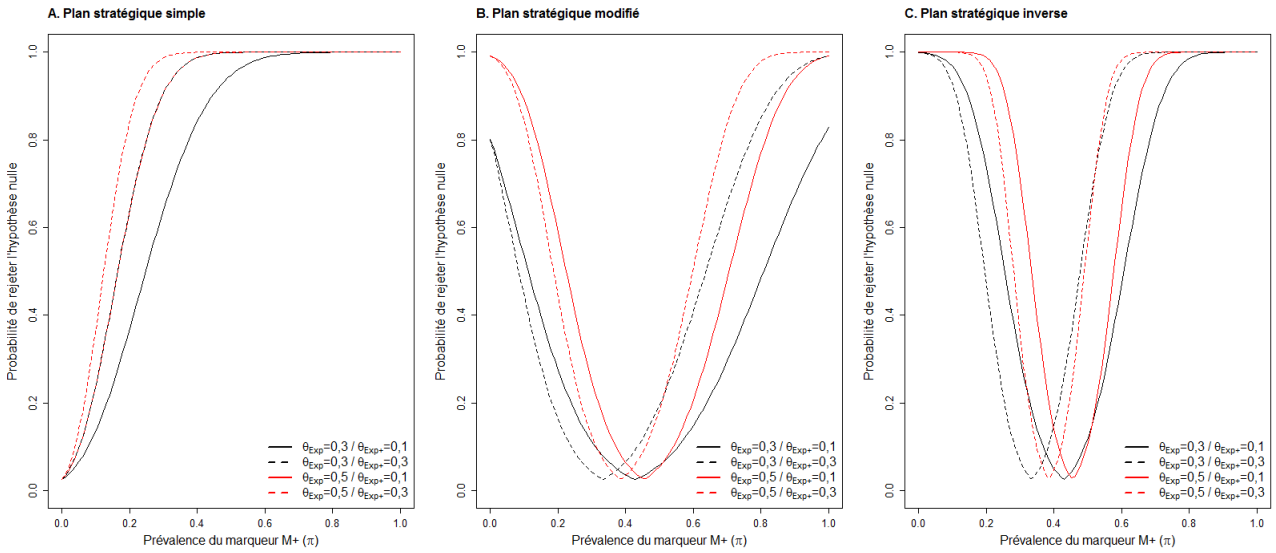


FIGURE 31 – Impact de la prévalence du marqueur $M+$ sur la probabilité de montrer une différence entre les deux stratégies lorsque le marqueur est prédictif de l'effet du traitement expérimental, pour différentes valeurs d'effet du traitement

Les graphes ont été obtenus avec les paramètres suivants : $\theta_0 = 0{,}1$, $\theta_{M+} = 0$, $\theta_{Std} = 0$ et $\theta_{Std+} = 0$ pour 100 patients dans chaque groupe de stratégie.

Les résultats obtenus sont similaires à ceux obtenus lorsque le marqueur n'est pas prédictif. Pour le plan expérimental stratégique simple, seuls des patients $M+$ peuvent bénéficier du traitement expérimental, et donc seule la valeur de p_{Exp+} a un impact sur la probabilité de rejeter H_0, quel que soit le ratio $\theta_{Exp}/\theta_{Exp+}$. Pour les plans expérimentaux stratégiques modifié et inverse, les valeurs de p_{Exp+} et de p_{Exp-} ont toutes deux un impact sur la probabilité de rejeter $H0$. Là encore H_0 n'est rejetée qu'à 5% pour $\pi = r$.

6.2.5 Discussion

L'identification et l'analyse des marqueurs prédictifs sont un moyen supplémentaire pour obtenir des informations concernant la réponse aux traitements. Cela peut soutenir le travail de découverte pré-clinique et accélérer le processus de découverte de traitements efficaces. L'utilisation d'une méthodologie d'analyse novatrice et statistiquement rigoureuse est essentielle à l'identification et à la validation des marqueurs prédictifs.

Les plans expérimentaux stratégiques permettent de valider l'utilité de mettre en place une médecine stratifiée pour le traitement expérimental. Il faut d'abord s'assurer du non effet bénéfique du traitement expérimental chez les patients $M-$. En effet, si le traitement

expérimental est aussi bénéfique pour les patients $M-$, l'interprétation des résultats de l'essai clinique utilisant un plan expérimental stratégique peut mettre en place, à tort, une médecine stratifiée. Un plan expérimental stratégique où le bras contrôle consisterait à donner le traitement expérimental à tous les patients (**plan stratégique avec contrôle expérimental**) permettrait de pallier ce problème (voir Tableau 8). Nos résultats confirment que les plans expérimentaux ne peuvent être mis en place qu'à condition d'avoir une forte preuve pré-clinique de la valeur prédictive du marqueur [118].

TABLE 8 – Interprétation des résultats d'un essai clinique utilisant un plan expérimental stratégique

Plan	Rejet de H_0 en faveur du bras stratégique	Rejet de H_0 en faveur du bras contrôle
Stratégique simple	Médecine stratifiée	Traitement standard uniquement
Stratégique modifié	Médecine stratifiée	Traitement standard ou expérimental
Stratégique inverse	Médecine stratifiée	Médecine stratifiée
Stratégique avec contrôle expérimental	Médecine stratifiée	Traitement expérimental uniquement

Dans le cas où le traitement expérimental est le meilleur pour tous les patients, seul le plan stratégique avec contrôle expérimental permet de conclure dans la bonne direction.

Lorsque le traitement expérimental ne peut pas être donné aux patients $M-$ pour des raisons éthiques de toxicité, le plan stratégique simple est le seul à pouvoir être employé. Sinon, les plans stratégiques modifié ou inverse doivent être mis en place, avec une préférence pour la stratégie inverse ayant une plus forte puissance que la stratégie modifiée, ce qui lui apporte un atout considérable dans le cas d'un faible effectif. En effet, dans notre calcul du nombre de sujets nécessaires pour les trois plans stratégiques, nous retrouvons les mêmes résultats que Eng [111] lorsque nous nous plaçons dans un cas de figure similaire à son étude, à savoir un effet du traitement expérimental meilleur chez les patients $M+$ et un effet du traitement standard meilleur chez les patients $M-$. Nous l'avons étendu à d'autres scénarios et avons montré que le plan expérimental stratégique inverse était toujours le moins couteux en nombre de sujets à inclure. Enfin, nous avons montré que dans le cas d'un petit échantillon, la valeur pronostique du marqueur impacte la puissance de l'étude et qu'il faut donc avoir les connaissances nécessaires sur la possibilité d'effet pronostique du marqueur avant la mis en place d'un tel plan expérimental.

Nous avons considéré le cas simple d'un marqueur prédictif binaire pour lequel le statut des patients était correctement identifié. Cependant, la puissance de l'étude peut diminuer lorsque la sensibilité et/ou la spécificité du test diagnostique associée(s) au

marqueur diminue(nt) [119, 116]. Nous souhaitons poursuivre notre étude en incluant les paramètres de sensibilité et de spécificité.

Notre étude est basée sur un critère de jugement binaire et un test de comparaison des proportions. Nous souhaitons poursuivre en considérant les critères de jugement continus ou les critères de survie. Lorsque le marqueur n'est pas issu d'une mesure binaire, mais quantitative, se pose aussi la question du *cut-off* utilisé pour discriminer les patients $M+$ et les patients $M-$. Si les traitements appropriés pour les patients $M+$ et $M-$ confirmés sont connus et validés, reste à s'assurer de donner le meilleur traitement aux cas intermédiaires. Dans ce cas des plans « hybrides » peuvent être mis en place où les patients intermédiaires sont les seuls à être randomisés entre les deux groupes de traitement [117].

Chapitre 7

Intégration des mesures cognitives grâce à la prise en compte de l'effet retest

7.1 Définition de l'effet retest et problématique associée

Soit un test construit pour mesurer les performances cognitives d'un patient au cours du temps. Soit Y_0 le score du patient à l'instant initial (T_0) et Y_1 le score du patient obtenu par exemple un an plus tard (T_1). Dans le cas de la maladie de Huntington, les performances du patient diminuent au cours du temps au fur et à mesure de la progression de la maladie ($Y_1 < Y_0$). Cependant, lorsque le patient passe le test pour la seconde fois, il a tendance à mieux appréhender le test et donc avoir de meilleurs résultats (figure 32).

Ainsi le résultat observé lors de la seconde passation est en fait une somme de deux effets opposés :

1. un déclin des performances dû à la progression de la maladie
2. une amélioration des performances due à la familiarité avec le test.

C'est ce deuxième point qui correspond à l'effet « retest » [53].

La familiarité au test provoquant l'effet retest est une somme de plusieurs facteurs : l'anticipation des questions, la mémorisation des réponses (notamment dans les tâches de mémorisation), la diminution du stress face à une situation connue, la mise en place d'une stratégie de réponse, ... Pour pallier ce problème, les tâches de mémorisation ont été élaborées avec des formes parallèles, c'est-à-dire avec différentes versions du même test évaluant le même critère tout en évitant le biais de mémorisation. Par exemple, dans le test de Hopkins consistant à mémoriser douze mots, il existe six formes parallèles. Pour pouvoir être utilisées, les formes parallèles doivent être équivalentes, c'est à dire que le

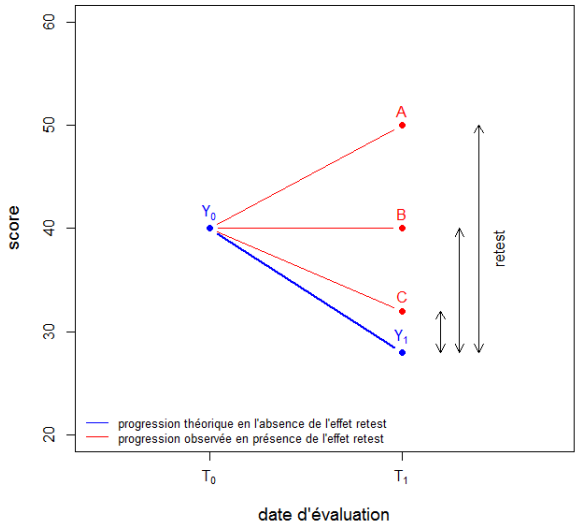


FIGURE 32 – Représentation schématique de l'effet « retest »

Y_0 représente le score obtenu lors de la première évaluation. Y_1 représente le score obtenu à la seconde évaluation en l'absence d'effet retest (déclin réel). En présence d'effet retest on peut observer (A) une amélioration des performances si le retest est supérieur au déclin, (B) une stabilité des performances si le retest compense le déclin et (C) un déclin inférieur au déclin réel si l'effet retest est petit comparé au déclin réel

patient doit obtenir le même score quelle que soit la forme avec laquelle il est évalué. Cela complexifie la validation de ce type de test, en particulier lorsque ces tests sont traduits dans une autre langue. Cependant, l'utilisation de formes parallèles ne permet pas de supprimer le biais lié à la familiarisation au test.

L'effet retest peut donc masquer le déclin réel des patients en améliorant artificiellement les performances lors de la seconde passation. Nous supposons que l'effet retest ne se produit qu'entre la première et la deuxième passation du test, et qu'il n'existe plus lors des évaluations suivantes.

7.2 Article « How to capitalize on the retest effect in future trials on Huntington's disease? »

Contexte

Evaluer le déclin cognitif des patients Huntington en un an avec uniquement deux mesures n'est donc pas envisageable si les échelles proposées sont soumises à un effet retest. Par conséquent, il faut adopter une stratégie qui permette de s'affranchir de l'effet

retest et de mesurer le déclin réel du patient entre T_0 et T_1. L'unique stratégie proposée pour le moment est de réaliser une double évaluation à T_0 [53]. Cette stratégie a été utilisée dans MIG-HD (voir Figure 2) ainsi que dans l'essai clinique évaluant le traitement Riluzole chez les patients Huntington (*Riluzole in Huntington's Disease*, RIL-HD) [120].

Méthode

Nous utilisons les données de ces deux essais cliniques afin de mesurer l'effet retest des échelles cognitives mais aussi des autres échelles utilisées dans la maladie de Huntington. Nous n'utilisons que les données de la première année de tous les patients de MIG-HD et les données de la première année du bras placebo de RIL-HD. Cela correspond à utiliser uniquement les données où les patients n'ont pas été traités afin de mesurer le retest et son impact sur la mesure de la progression naturelle de la maladie. Pour les deux essais cliniques, les patients ont été évalués à T_0, $T_0 + \Delta$ et T_1 où T_0 est la première évaluation, $T_0 + \Delta$ est une seconde évaluation où $\Delta = 1$ mois pour MIG-HD et $\Delta = 2$ semaines pour RIL-HD et où T_1 est une troisième évaluation un an plus tard. Nous comparons les scores obtenus à T_0 et à $T_0 + \Delta$ pour savoir si les patients sont effectivement meilleurs à la seconde passation du test sur un intervalle assez court, pour lequel il n'y a pas de déclin. De plus, nous vérifions si utiliser $T_0 + \Delta$ comme première évaluation de référence à la place de T_0 permet de mesurer un déclin qui n'est pas mesurable en utilisant T_0 uniquement. De plus, pour chaque échelle, nous avons développé un modèle linéaire permettant de prédire les performances des patients à un an.

Discussion

Ce travail nous a permis de montrer l'existence de l'effet retest dans les tests cognitifs. Les résultats obtenus sur les scores moteurs et fonctionnels montrent qu'ils ne sont pas soumis à l'effet retest, ce qui conforte le choix de leur utilisation dans les essais cliniques.

Cependant, dans une maladie neurodégénérative, le critère cognitif est tout aussi important. En effet, le déclin cognitif a un impact dans la vie de tous les jours notamment dans la relation du patient avec son entourage, dans son travail, ...

Nous avons montré que grâce à la stratégie consistant à réaliser une double évaluation, certains tests cognitifs peuvent montrer un déclin. Nous suggérons donc que dans les prochains essais cliniques longitudinaux portant sur la maladie de Huntington, soit mise en place une double évaluation à l'inclusion dès qu'un critère (principal ou secondaire) est cognitif. Dans ce cas, seule la seconde mesure doit être utilisée dans l'analyse longitudinale.

Cependant, la stratégie de double évaluation repose sur l'hypothèse d'un effet retest uniquement présent entre la première et la seconde passation du test. Nous considérons donc que les conditions de l'évaluation à T_1 sont comparables à celles de l'évaluation à

$T_{0,bis} = T_0 + \Delta$. Or, on pourrait se demander si l'effet retest sur une courte période Δ pourrait exister lors d'une seconde double évaluation à T_1 et $T_{1,bis} = T_1 + \Delta$, et comment il est impacté par Δ et par le laps de temps séparant les doubles évaluations (T_0,$T_{0,bis}$) et (T_1,$T_{1,bis}$). Cette question est importante pour les futurs essais cliniques afin de savoir à quelle fréquence les patients doivent être testés et si une double évaluation est nécessaire uniquement au début ou tout au long du protocole. La figure 33 propose plusieurs scénarios et ce qu'ils impliquent pour un essai clinique.

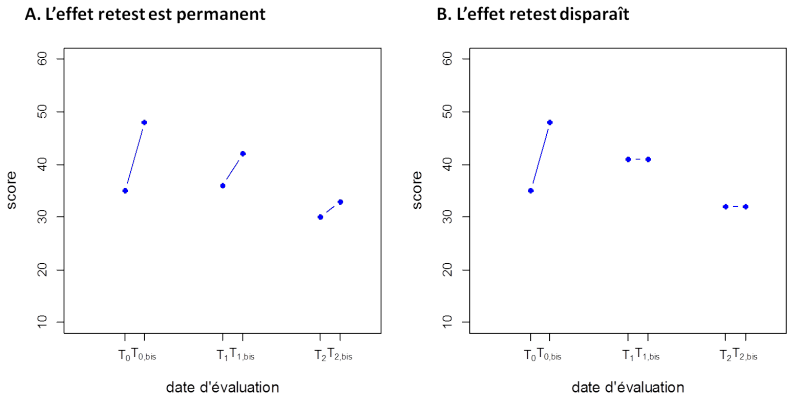


FIGURE 33 – Représentation schématique d'un effet « retest » permanent ou non permanent

(A) Le patient s'améliore entre deux évaluations consécutives séparées par un faible intervalle de temps, mais cette amélioration s'atténue au cours du temps. (B) Le patient s'améliore entre la première et la seconde évaluation puis l'effet retest disparaît lors des évaluations suivantes.

D'autre part, dans le cas où un effet retest sur une période Δ est permanent, il serait intéressant de modéliser l'évolution du retest pour savoir s'il s'atténue au fil du temps. Probablement, le retest diminuera chez les patients atteints de la maladie de Huntington. En effet, le retest correspond à un apprentissage. Plus la maladie progresse, plus l'apprentissage sera compliqué. Mais s'il ne diminue pas chez des sujets contrôles, le retest pourrait être un marqueur du déclin cognitif. Ainsi si le retest persévère au cours du temps, connaître l'impact du déclin cognitif sur le retest est aussi une piste pour l'évaluation d'une facette du déclin cognitif.

PLOS ONE

RESEARCH ARTICLE

How to Capitalize on the Retest Effect in Future Trials on Huntington's Disease

Catherine Schramm[1,2,3,4], Sandrine Katsahian[2,5], Katia Youssov[1,3,4,6], Jean-François Démonet[7], Pierre Krystkowiak[8,9,10], Frédéric Supiot[11], Christophe Verny[12], Laurent Cleret de Langavant[1,3,4,6], Anne-Catherine Bachoud-Lévi[1,3,4,6]*, European Huntington's Disease Initiative Study Group and the Multicentre Intracerebral Grafting in Huntington's Disease Group[¶]

1 INSERM U955 E01, Neuropsychologie interventionnelle, Institut Mondor de Recherche Biomédicale, Créteil, France, **2** INSERM UMRS1138 E22, Science de l'information au service de la médecine personnalisée, Centre de Recherche des Cordeliers, Université Paris 5, Université Paris 6, Paris, France, **3** Université Paris Est, Faculté de Médecine, Créteil, France, **4** Ecole Normale Supérieure, Institut d'Etude de la Cognition, Paris, France, **5** Assistance Publique-Hôpitaux de Paris, Service d'informatique et statistiques, Hôpital Européen Georges Pompidou, Paris, France, **6** Assistance Publique-Hôpitaux de Paris, Centre National de Référence pour la Maladie de Huntington, Hôpital Henri Mondor, Créteil, France, **7** Leenaards Memory Centre, Clinical Neurosciences Department, CHUV Lausanne, Lausanne, Switzerland, **8** Centre Hospitalier Universitaire d'Amiens, Service de neurologie, Amiens, France, **9** EA 4559 - Laboratoire de Neurosciences Fonctionnelles et Pathologie (LNFP), Université de Picardie Jules Verne (UPJV), Amiens, France, **10** SFR CAP-Santé (FED 4231), Amiens, France, **11** Hôpital Erasme ULB, Service de Neurologie, Bruxelles, Belgium, **12** CHU d'Angers, Centre de Référence des Maladies Neurogénétiques, Service de Neurologie, Angers, France

¶ Membership of the European Huntington's Disease Initiative Study Group and the Multicentre Intracerebral Grafting in Huntington's Disease Group is provided in the Acknowledgments.
* bachoud@gmail.com

OPEN ACCESS

Citation: Schramm C, Katsahian S, Youssov K, Démonet J-F, Krystkowiak P, Supiot F, et al. (2015) How to Capitalize on the Retest Effect in Future Trials on Huntington's Disease. PLoS ONE 10(12): e0145842. doi:10.1371/journal.pone.0145842

Editor: David Blum, Inserm U837, FRANCE

Received: July 7, 2015

Accepted: December 9, 2015

Published: December 29, 2015

Data Availability Statement: The consent form specifies that our institution guarantees the data confidentiality. We thus can provide the data individually upon request, while discarding any potential identifying data and after inquirer's agreement for both use and protection of the data. Please contact Professor AC Bachoud-Lévi (anne-catherine.bachoud-levi@aphp.fr).

Funding: This study was supported by investment for the future NeurATRIS: Infrastructure de recherche translationnelle pour les biothérapies en Neurosciences (ANR-11-INBS-0011, http://www.agence-nationale-recherche.fr/), European Community Seventh Framework Program

Abstract

The retest effect—improvement of performance on second exposure to a task—may impede the detection of cognitive decline in clinical trials for neurodegenerative diseases. We assessed the impact of the retest effect in Huntington's disease trials, and investigated its possible neutralization. We enrolled 54 patients in the Multicentric Intracerebral Grafting in Huntington's Disease (MIG-HD) trial and 39 in the placebo arm of the Riluzole trial in Huntington's Disease (RIL-HD). All were assessed with the Unified Huntington's Disease Rating Scale (UHDRS) plus additional cognitive tasks at baseline (A_1), shortly after baseline (A_2) and one year later (A_3). We used paired *t*-tests to analyze the retest effect between A_1 and A_2. For each task of the MIG-HD study, we used a stepwise algorithm to design models predictive of patient performance at A_3, which we applied to the RIL-HD trial for external validation. We observed a retest effect in most cognitive tasks. A decline in performance at one year was detected in 3 of the 15 cognitive tasks with A_1 as the baseline, and 9 of the 15 cognitive tasks with A_2 as the baseline. We also included the retest effect in performance modeling and showed that it facilitated performance prediction one year later for 14 of the 15 cognitive tasks. The retest effect may mask cognitive decline in patients with neurodegenerative diseases. The dual baseline can improve clinical trial design, and better prediction should homogenize patient groups, resulting in smaller numbers of participants being required.

Neurostemcell (Grant Agreement no. 222943, http://ec.europa.eu/research/fp7/), European Community Seventh Framework Program Repair-HD (Grant Agreement no 602245, http://ec.europa.eu/research/fp7/). The Département d'Etudes Cognitives of the Ecole Normale Supérieure is supported by two ANR grants from the French Research Agency (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL, http://www.agence-nationale-recherche.fr/). Assistance Publique-Hôpitaux de Paris is the sponsor for the MIG-HD study (Ref NCT00190450) and Sanofi Aventis for the Riluzole study (Ref NCT00277602). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: CS was successively supported by the NeuroStemcell Consortium (European Community Seventh Framework Program grant agreement no. 222943) and by "Investments for the future" (ANR11INBS0011 NeurATRIS: Infrastructure de recherche translationnelle pour les biothérapies en Neurosciences). SK: no financial disclosures. KY: no financial disclosures. JFD has received financial support from Eli Lilly, Lundbeck, Novartis, Schwabe and Vifor Pharma over the Rebuttal letter past 2 years as a member of scientific boards and speaker at sponsored sessions. This financial support was completely unrelated to the work reported here. PK: no financial disclosures. FS: no financial disclosures. CV: no financial disclosures. LCL: no financial disclosures. CBL acted as a consultant for Teva, once, in 2014. She received grants from the Ministry of Health supporting the National Reference Center for Huntington's Disease and several grants for academic trials provided by the Direction de la Recherche Clinique (APHP). She is a partner in several investments for the future projects (Labex IEC, Neuratris) and in an EU FP7 project (RepairHD). This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

Introduction

Huntington's disease (HD) is an inherited neurodegenerative disorder involving motor, behavioral and cognitive impairments [1]. The cognitive disorders have a major impact on daily life, but most clinical trials focus on motor endpoints. This is because clinical trial endpoints must be able to capture both patient decline and treatment efficacy, and cognitive decline is much more difficult to capture within one year in patients at early disease stages [2] than motor decline. This difficulty of assessment results from the heterogeneity of cognitive changes (language, memory, etc.) and two opposing effects: the retest effect and patient decline due to disease progression. The retest effect is defined as an improvement in performance with repeated exposure to a task, with the greatest improvement occurring between the first two assessments [3–5]. This effect combines familiarity with the task and testing environment and the possible recall of responses [2]. The first assessment, during which everything is new to the patient, is always the most difficult.

The retest effect may have contributed to the failure of some neuroprotection trials, by adding noise to statistics comparing patients with different backgrounds at baseline, particularly in trials including small numbers of patients, such as those assessing biotherapy. One approach to neutralizing the retest effect is to carry out a second assessment (A_2) shortly after the first (A_1), and then discard the results obtained at A_1 from the analysis, using performance at A_2 as the baseline [2]. In addition, the retest effect (ΔA_2-A_1) can be used to improve the prediction of long-term patient performance. Indeed, in an observational longitudinal study in HD patients, the retest effect (ΔA_2-A_1 around 7 months) accounted for up to 36% of the variance of performance at A_3 (ΔA_3-A_2 around 29 months) [6]. Likewise, in healthy elderly adults, performance at A_3 (one year) is accurately predicted by the one week-interval retest effect (ΔA_2-A_1) [7].

However, the impact of the retest effect in clinical trials, which include additional variability (placebo effect, hope, anxiety about treatment and randomization), remains unknown. Two trials, the *Multicentric Intracerebral Grafting in Huntington's Disease* (MIG-HD) [8] and *Riluzole in Huntington's Disease* (RIL-HD) [9] trials, were designed with a short-term test-retest procedure. We used the MIG-HD trial (i) to assess whether the retest effect modified performance and whether our strategy of using the second assessment as a baseline was sensitive to cognitive decline in the long-term (A_3) and (ii) to evaluate whether introducing the retest effect (ΔA_2-A_1) into the model of disease progression in patients improved the predictive value of the model in the long term (A_3). Finally, we transferred the models obtained for the MIG-HD cohort to the RIL-HD cohort, to assess their predictive value in another population.

Materials and Methods

Participants and design

Patients were enrolled in two separate trials: the MIG-HD trial ($N = 54$, Ref NCT00190450, PI AC Bachoud-Lévi) [8], which is currently underway, and the placebo group of the cognitive ancillary study of the RIL-HD trial conducted only in France ($N = 39$, Ref NCT00277602, study coordinator Sanofi) [9]. Both trials were approved by the institutional review board (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale) of Henri-Mondor Hospital at Créteil (MIG-HD the September 25, 2001, and RIL-HD the December 18, 2002). Patients had signed an inform consent. The data were analyzed anonymously.

The MIG-HD trial is a phase II randomized trial assessing the efficacy of cell transplantation in HD patients at early stages of the disease. Patients were assessed at inclusion (A_1), then 35 days (SD = 15) later (A_2). They were randomized at one year (A_3), to determine the timing of

transplantation (M_{13}-M_{14} for the early graft group or M_{33}-M_{34} for the late graft group). Patients were followed up until 52 months.

The RIL-HD trial is a phase III multinational, randomized, placebo-controlled, double-blind study, for which a cognitive ancillary study was conducted in France from 1999 to 2004, on patients with moderately advanced HD. Patients were assessed at inclusion (A_1), 15 days (SD = 8) later (A_2) and at one year (A_3), with randomization at A_2.

The demographic features for patients at A_1 are displayed in Table 1.

Clinical assessments

The Unified Huntington's Disease Rating Scale (UHDRS) [10] and additional cognitive tests were used in both studies. Motor score reflected both voluntary and involuntary capacity and ranged from 0 to 124 (highest severity). Functional disability was assessed with Total Functional Capacity (TFC, range: 13 to 0) and Independence Scale (IS, range 100 to 0) scores, with lower scores indicating greater functional impairment, and the Functional Assessment Scale (FAS, 25 to 50), with higher scores indicating greater functional impairment. The severity and frequency of behavioral dysfunctions were quantified with the behavioral part of the UHDRS (range: 0 to 88), with higher scores indicating greater impairment. Global cognitive efficiency was evaluated with the Mattis Dementia Rating Scale (MDRS) [11]. Several tasks were used to assess attention and executive functions: letter fluency (for P, R and V in French) determined for 1 minute, the Symbol Digit Modalities Test (SDMT), the three components of the Stroop test (color naming, word reading, and color-word interference), each assessed for 45 seconds [12], categorical fluency (for animals) assessed for 1 minute [13],[14], the Trail-Making Test forms A and B (TMT A and B) [15], scoring the time taken to link 25 points, with a maximal time of 240 seconds, and figure cancellation tasks [16], in which patients were asked to cross out one, two and then three figures from a panel of signs, in 90 seconds, with lower scores indicating greater cognitive impairment. Short-term and long-term memory were evaluated with the Hopkins Verbal Learning Task (HVLT) including immediate recall, delayed recall and recognition tasks [17],[18]. By contrast to the other tasks, the HVLT was assessed with alternating parallel forms.

Each patient performed one motor test, three functional tests, one behavioral test and 15 cognitive tests at each assessment point.

Table 1. Characteristics of patients at their inclusion (A_1) in the MIG-HD and RIL-HD trials.

Characteristics	MIG-HD (*N* = 54)	RIL-HD (*N* = 39)
Age, y, mean (SD)	43.3 (8.7)	48.5 (10.1)
Sex % men / women	63.0 / 37.0	48.7 / 51.3
Education level, y, mean (SD)	12.0 (3.4)	12.3 (3.6)
Inheritance % paternal / maternal	60.0 / 40.0	47.6 / 52.4
Age of parent at onset, y, mean (SD)	42.2 (10.6)	45.7 (10.8)
Number of CAG repeats, mean (SD)	45.4 (4.2)	44.1 (3.6)
Time since onset, y, mean (SD)	4.5 (2.6)	6.1 (6.2)
TFC, mean (SD)	11.7 (1.0)	10.8 (1.8)
First symptom %		
Motor	60.7	70.3
Cognitive	17.9	13.5
Psychiatric	21.4	16.2

y: years; SD: standard deviation; TFC: total functional capacity.

doi:10.1371/journal.pone.0145842.t001

Statistical Analysis

Evaluation of the retest effect in the MIG-HD cohort. For each task, we used Student's *t*-tests for paired data to compare performances, first between A_1 and A_2, to measure the potential retest effect, then between A_1 and A_3, to assess the decline over a one-year period and between A_2 and A_3, to determine whether discarding the A_1 data unmasked a decline that was otherwise undetectable.

Modeling of performance for the MIG-HD cohort. For each task, we selected the multivariate linear model best predicting the data at one year, by stepwise selection [19] with the Akaike Information Criterion (AIC) [20]. We used an iterative algorithm (stepwise selection) to select, without prior assumptions, the best predictive factors from a set of 10 variables (performance at A_1, retest, age, sex, education level expressed as the number of years spent studying, parental inheritance, age of parent at disease onset, CAG repeat length, time since disease onset and the nature of the first symptom appearing at disease onset (motor, cognitive or psychiatric), as determined by the clinician or, if no clinician's assessment was available, by the family or the patient). Lower AIC values indicate a better fit of the model to the data. The first model selection step was carried out for patients with complete data sets only. Estimates of regression coefficients were refined, by recalculating each model, using all the available complete data for the selected variables. The retest is the difference: performance at A_2 –performance at A_1 and is denoted ΔA_2-A_1. For each task, performance at A_3 (P) was predicted as follows:

$$P = \beta_0 + \beta_{score\ at\ A_1} \times \text{performance at } A_1 + \beta_{retest} \times \Delta A_2 - A_1 + \beta_{age\ at\ A_1} \times \text{age} + \beta_{sex} + \beta_{education\ level} \times \text{education level} + \beta_{inheritance} + \beta_{age\ of\ parent\ at\ onset} \times \text{age of parent at onset} + \beta_{CAG} \times \text{CAG} + \beta_{time\ since\ onset} \times \text{time since onset} + \beta_{first\ symptom}$$

where age, education level and age of parent at onset are expressed in years; the first symptom could be motor, cognitive or psychiatric; β_0 is the intercept and, for each variable, $\beta_{variable}$ is its associated regression coefficient (0 for the variables not selected). For quantitative variables, $\beta_{variable}$ was multiplied by the value of the variable. For qualitative variables (sex, inheritance and first symptom), "woman", "maternal inheritance" and "motor symptom" constituted the reference factors, such that $\beta_{woman} = \beta_{maternal} = \beta_{motor} = 0$. Calculation of the associated 95% predictive interval (95% PI) is explained in the supplemental data (S1 Text).

External validation on the RIL-HD cohort. We used models constructed from data for the MIG-HD cohort to predict performances at A_3 for each patient in the RIL-HD cohort. Then, for each task, we measured the concordance between observed (O) and predicted (P) values, using the intraclass correlation coefficient (ICC) and the coefficient of determination (R_e^2). The ICC was calculated with a two-way mixed effect model [21] and evaluates agreement between observed (O) and predicted (P) performances at A_3 in the RIL-HD cohort. The coefficient of determination (R_e^2) is the percentage of the observed performance variance explained by the model constructed from MIG-HD data. It assesses the degree to which observed performance at A_3 in the RIL-HD cohort is accurately predicted by the model, as follows:

$$R_e^2 = 1 - \frac{\sum_i (O_i - P_i)}{\sum_i (O_i - m)}$$

where i refers to a patient and m is the mean observed performance at A_3. $R_e^2 = 1$ indicates a

perfect predictive value of the model, whereas $R_e^2 \leq 0$ indicates that the model is not informative.

Analyses were performed with R 2.13 software (http://www.r-project.org/). All tests were two-tailed and values of $P < 0.05$ were considered significant.

Results

Evaluating the retest effect in the MIG-HD cohort

We assessed the retest effect between A_1 and A_2 in the MIG-HD cohort. Performance improved in seven cognitive tasks, and remained stable in the other cognitive, motor and functional tasks, except for FAS score, which declined between A_1 and A_2 (Fig 1).

We assessed decline between A_1 and A_3 and between A_2 and A_3 in the MIG-HD cohort (Fig 2). The use of A_2 as the baseline increased the number of tasks for which a decline in performance was detected from three to nine, but FAS score was the only motor or functional performance affected. Indeed, FAS performance declined between A_1 and A_3 but not between A_2 and A_3. Behavioral performance improved between A_2 and A_3.

Modeling of performance in the MIG-HD cohort

Table 2 displays the regression coefficients of the predictive model for each task, for the MIG-HD cohort. Performance at A_1 was predictive of performance at A_3 in all tasks. Introducing the difference in performance between A_1 and A_2 (ΔA_2-A_1) into the models improved the prediction of performance at A_3 for 14 of the 15 cognitive tasks, for behavioral and motor performance and TFC. Larger numbers of CAG repeats were associated with a poorer FAS and IS scale scores and poorer motor performance, but better behavioral performance. Women outperformed men in 7 of the 15 cognitive tasks. Sex had no effect on motor and functional performances, whereas behavioral performance was better in women than in men. Higher education levels were associated with better performance at A_3 for all components of the HVLT.

The regression coefficients presented in Table 2 are those used in the predictive models. For example, the performance at A_3 in letter Fluency 1' is given by the following formula:

$$\text{performance at } A_3 = \begin{cases} 10.27 + 0.66 \times \text{performance at } A_1 + 0.84 \times \text{retest} & \text{woman} \\ 10.27 + 0.66 \times \text{performance at } A_1 + 0.84 \times \text{retest} - 2.55 & \text{man} \end{cases}$$

The equations associated with the predictive models for each task are detailed in S1 Table. Moreover, S2 Table gives additional parameters for calculation of the 95% PI.

External validation on the RIL-HD cohort

For each task, we determined the predictive value of models by calculating the ICC and R_e^2 (Fig 3). Performance in the RIL-HD trial was well predicted for 14 of 20 tasks by the models developed with data for the MIG-HD cohort ($R_e^2 \geq 0.5$ and ICC ≥ 0.6).

Discussion

The design of clinical trials for neurodegenerative diseases could be improved by methodological approaches based on our knowledge of the patient's cognitive performances. However, cognitive knowledge is obtained mostly through longitudinal follow-up in observational studies, which may not include variability factors inherent to clinical trials. The retest effect may impede observations of cognitive decline in patients with Huntington's disease. We therefore assessed its impact in two long-term clinical trials in HD patients, with a short interval between

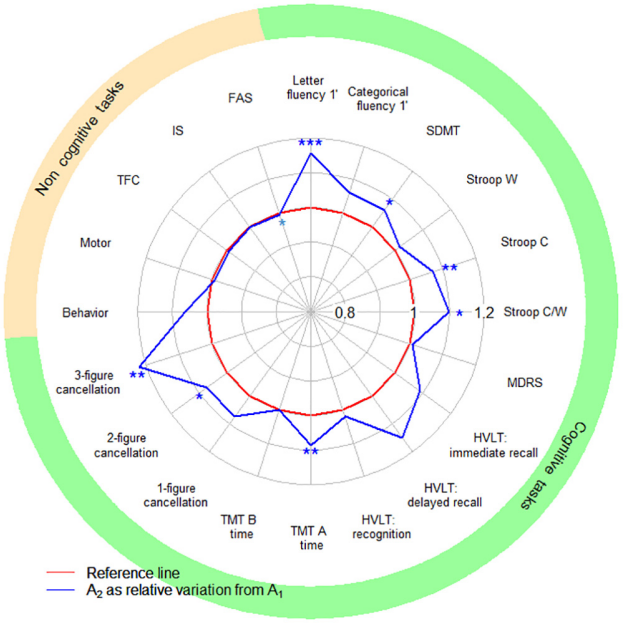


Fig 1. Impact of the retest effect in the MIG-HD cohort. SDMT: Symbol Digit Modalities Test; Stroop C, W and C/W: Stroop color, word and color/word interference; MDRS: Mattis Dementia Rating Scale; TMT A, B: Trail-Making Test A and B; TFC: Total Functional Capacity; IS: Independence Scale; FAS: Functional Assessment Scale. The red curve represents the baseline (reference score A_1) and the blue curve shows the mean relative score one month later (A_2). The portion of the blue curve beyond the red curve indicates performance improvement between A_1 and A_2. Paired *t*-tests, significance: * $P<0.05$, ** $P<0.01$, *** $P<0.001$.

doi:10.1371/journal.pone.0145842.g001

the first and second assessments (MIG-HD, RILH-HD). We first determined whether there was a detectable retest effect between the first two assessments (A_1 and A_2), and then evaluated the impact of this effect one year later (A_3). We found that replacing A_1 with A_2 as the baseline unmasked a decline that would not otherwise have been detected. Indeed, the comparison between A_2 and A_3 showed declines that were not apparent in the comparison between A_1 and A_3. We also modeled patient performance and showed how the inclusion of the retest effect in patient performance models would improve trial design.

At one year, decline was observed in a few cognitive tasks (SDMT, MDRS and the HVLT immediate recall), the motor task and all functional tasks. However, consistent with previous findings [2], there was a pronounced retest effect in cognitive tasks (letter fluency, SDMT, Stroop color and color/word interference, TMT A and 2- and 3-figure cancellation tasks), but no such effect in motor and functional assessments. This retest effect may hamper the objective detection of cognitive decline, with a major impact in tasks with a high cognitive demand, obscuring performance decline over a one-year period [22]. Neutralization of the retest effect is particularly important in clinical trials, because some patients may already have been exposed to testing whereas others have not, adding background noise to the overall performance data. Assuming that the retest effect is maximal at the second assessment, the use of this assessment as the baseline can decrease the impact of the retest effect on subsequent assessments. By discarding performances at A_1 and using the performance measured at A_2 as the baseline, we unmasked a decline in six tasks (Stroop color and color/word interference, recognition part of HVLT, TMT A and 2- and 3-figure cancellation), demonstrating the efficacy of this strategy for small samples. However, the improvement in behavioral performance [23], contrasting

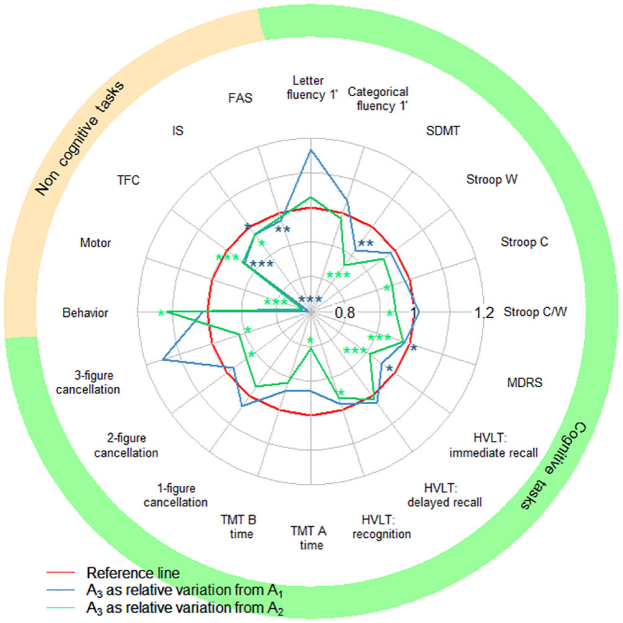


Fig 2. Observed performance at one year (A_3), with A_1 or A_2 used as the baseline, in the MIG-HD cohort. SDMT: Symbol Digit Modalities Test; Stroop C, W and C/W: Stroop color, word and color/word interference; MDRS: Mattis Dementia Rating Scale; TMT A, B: Trail-Making Test A and B; TFC: Total Functional Capacity; IS: Independence Scale; FAS: Functional Assessment Scale. The red curve represents the baseline (reference score). The blue (or green) curve corresponds to the mean relative score one year later (A_3), with A_1 (or A_2 for the green curve) used as the baseline. A green curve within the blue curve indicates that the decline was easier to detect if A_2 was used as the baseline, rather than A_1. Paired *t*-tests, significance: * $P<0.05$, ** $P<0.01$, *** $P<0.001$.

doi:10.1371/journal.pone.0145842.g002

with the decline in other task performances, may reflect the patients' hopes and expectations of treatment.

The HVLT constitutes a specific case: we alternated parallel forms because of the strength of item recall in declarative memory tasks [24]. However, alternation was not used for other tasks, because parallel forms are of no interest for procedural tasks or tasks with a strong motor output (SDMT, TMT A and verbal fluency tasks) [25]. The use of parallel forms should also be limited because of their low intrasubject equivalence, potentially introducing noise into longitudinal performance analyses. Furthermore, the ceiling effect observed in patients with high scores in the HVLT, MDRS and TMT tasks limits the utility of neutralizing the retest effect.

However, the retest effect depends not only on the nature of the task, but also on the population assessed [26]. Indeed, Cooper *et al.* [27], [28] demonstrated the existence of a retest effect in categorical fluency assessment in healthy participants but not in patients with Alzheimer's disease or mild cognitive impairments. Likewise, we found no retest effect for this task in HD patients.

In addition to masking decline, the retest effect may provide information about disease progression [7]. This suggests that combining a strategy based on the individual performance of patients and the nature of the tasks may be useful. Indeed, the modeling of patient performance at one year for each task showed that ΔA_2-A_1 performance, even in the absence of a significant retest effect, accurately predicted performance for most cognitive tasks in HD and for motor and behavior tasks and TFC. ΔA_2-A_1 performance appears to be more frequently selected by stepwise algorithms than sociodemographic and genetic variables. We also arbitrated between parameters to strengthen our models. For example, both the number of CAG repeats and age

Table 2. Predictive factors for each task.

		β_0	$\beta_{score\ at\ A1}$	β_{retest}	$\beta_{age\ at\ A1}$	$B_{sex\ =\ man}$	$B_{education\ level}$	$\beta_{inheritance\ =\ paternal}$	$\beta_{age\ of\ parent\ at\ onset}$	β_{CAG}	$B_{time\ since\ onset}$	$\beta_{first\ symptom\ =\ cognitive}$	$\beta_{first\ symptom\ =\ psychiatric}$
Cognitive	**Letter Fluency 1'**	10.27* (3.92)	0.66*** (0.14)	0.84*** (0.18)		-2.55 (2.51)							
	Categorical Fluency 1'	6.55** (1.98)	0.57*** (0.13)	0.55*** (0.15)		-1.82 (0.93)							
	SDMT	-0.84 (2.19)	0.98*** (0.07)	0.33* (0.15)		-1.93 (1.22)							
	Stroop W	1.56 (8.67)	0.93*** (0.13)	1.04*** (0.22)									
	Stroop C	3.03 (5.66)	1.01*** (0.10)	0.43* (0.18)		-8.43** (2.58)							
	Stroop W/C	2.07 (3.17)	0.97*** (0.10)	0.65*** (0.14)		-3.65* (1.64)							
	HVLT: Immediate recall	6.08 (3.07)	0.53*** (0.10)	0.27* (0.12)		-2.01 (1.11)	0.26 (0.17)					-2.08 (1.37)	0.45 (1.22)
	HVLT: delayed recall	-0.51 (1.18)	0.55*** (0.10)				0.23* (0.09)						
	HVLT: recognition	-1.35 (1.60)	0.87*** (0.14)	0.52*** (0.13)			0.19** (0.06)						
	MDRS	20.29 (13.26)	0.89*** (0.09)	0.64*** (0.14)	-0.10 (0.07)	-2.81* (1.25)		-1.12 (1.22)			-0.39 (0.25)	0.18 (1.63)	1.97 (1.32)
	1-figure cancellation	1.51 (1.54)	0.89*** (0.08)	0.57*** (0.14)									
	2-figure cancellation	0.24 (1.55)	0.93*** (0.08)	0.50** (0.16)									
	3-figure cancellation	7.28* (2.74)	0.83*** (0.11)	0.55* (0.20)									
	TMT A time	13.11 (10.81)	0.90*** (0.15)	0.59** (0.18)									
	TMT B time	28.68* (22.40)	0.94*** (0.06)	0.86*** (0.11)	-0.98 (0.56)				0.63 (0.44)				
	Behavior	40.26** (13.96)	0.31* (0.12)	0.52** (0.16)		2.91 (2.08)	-0.92** (0.32)			-0.68* (0.28)	1.36** (0.48)		
	Motor	-32.91 (26.47)	0.81*** (0.1)	0.68** (0.21)	0.52* (0.25)			4.40 (2.67)	-0.23 (0.14)	0.63 (0.43)			
Functional	**FAS**	0.42 (5.44)	0.65*** (0.14)		0.03 (0.04)					0.18* (0.07)			
	IS	70.41* (28.28)	0.63*** (0.16)		-0.22 (0.19)					-0.66 (0.38)			
	TFC	-0.55 (2.18)	0.98*** (0.19)	1.50*** (0.4)	-0.003 (0.02)			0.35 (0.34)					

SDMT: Symbol Digit Modalities Test; Stroop C, W and C/W: Stroop color, word and color/word interference; HVLT: Hopkins Verbal Learning Task; MDRS: Mattis Dementia Rating Scale; TMT A, B: Trail-Making Test A and B; FAS: Functional Assessment Scale; IS: Independence Scale; TFC: Total Functional Capacity. A given row shows the predictive factors (estimated regression coefficient, standard error and significance: * $P<0.05$, ** $P<0.01$, *** $P<0.001$) for the corresponding task. The absence of a value indicates that the covariate concerned was not selected for the model.

doi:10.1371/journal.pone.0145842.t002

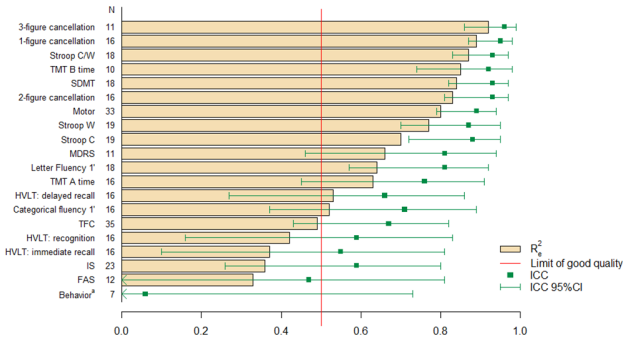


Fig 3. External validation of models in the RIL-HD cohort, based on R_e^2 and ICC. SDMT: Symbol Digit Modalities Test; Stroop C, W and C/W: Stroop color, word and color/word interference; MDRS: Mattis Dementia Rating Scale; TMT A, B: Trail-Making Test A and B; HVLT: Hopkins Verbal Learning Task; TFC: Total Functional Capacity; IS: Independence Scale; FAS: Functional Assessment Scale. N: number of patients in the RIL-HD cohort for whom all the data required for the predictive model were available. R_e^2: coefficient of determination for external validation. ICC: intraclass correlation coefficient. 95% CI: 95% confidence interval. a: R_e^2 = -0.7. The red line represents the limit for a high-quality model (R_e^2 > 50% of the observed variance explained by the model).

doi:10.1371/journal.pone.0145842.g003

at onset are eligible variables [29], but they are correlated [30–32], so only one of these factors should be included in the model [33]. We decided to include the number of CAG repeats, as age at onset is subject to some degree of subjectivity. Likewise, rather than using the performance in one task to explain performance in another task (e.g. using motor score to explain TFC [34]), we limited the set of eligible variables to demographic variables. Finally, we did not include handedness in our models, because 90% of the patients were right-handed.

This approach made it possible to include a larger number of covariates in our models than in those of previous studies and to prioritize them through the selection algorithm. For example, the number of CAG repeats has been reported to affect general verbal and spatial abilities [35], whereas our stepwise selection suggested that it was predictive of performance in the 3-figure cancellation task, which has a spatial nonverbal component. Indeed, the number of CAG repeats was found to have less impact than the sex of the patient in verbal tasks (letter and categorical fluencies) and sex was not included in the model described in the previous study. Furthermore, dichotomization of the number of CAG repeats variable (small and large numbers of repeats) may have resulted in greater importance being assigned to this variable than in models, such as ours, in which the number of CAG repeats was treated as a continuous variable. Like Ruocco [36], Kieburtz [37] and Feigin *et al.* [38], we showed that the number of CAG repeats improved in the prediction of motor performance, but not TFC. Finally, higher education levels were associated with a better performance, for all HVLT components.

The small number of patients enrolled in the MIG-HD study is a potential limitation in the search for predictive factors for future studies. However, external validation on the RIL-HD cohort, through calculation of the intraclass correlation coefficient and the determination coefficient (R_e^2), demonstrated the reproducibility and robustness of our models, regardless of the differences between the two trials. Indeed, patients in the MIG-HD trial were not randomized until one year (A_3), whereas those in the RIL-HD study were randomized at the second assessment (A_2). Consequently, the patients in the MIG-HD study approached the intervention with greater hope, whereas those in the placebo group of the RIL-HD study may have been aware of a lack of improvement during the follow-up period. This difference may account for the poor prediction of behavioral performance in the RIL-HD study ($R_e^2 < 0$). By contrast, the difference in time interval between A_1 and A_2 in the two studies had no impact on prediction quality,

further demonstrating the validity of the models. The models were constructed with data from patients with relatively mild disease. They may, therefore, not be applicable to patients with more advanced HD. Indeed, retest effects would be expected to be smaller in patients with more severe disease.

Our findings indicate that the retest effect is a limitation in clinical trials, but that both its neutralization, through the use of a second assessment as a baseline, and its integration into task modeling would be beneficial in future trials. For example, our predictive models may facilitate the identification of rapid decliners [39], defined as individuals whose observed performance is worse than predicted. Indeed, in longitudinal clinical trials, treatment effects could be masked in such patients, as already shown for Alzheimer's disease [40]. The identification of such patients is helpful for trial design, in two ways. First, the exclusion of such patients would probably decrease intersubject variability, making it possible to decrease sample size. Second, rapid decliners could be uniformly allocated to the different arms of the study by stratified randomization, to ensure the constitution of comparable groups, in terms of both baseline data and disease progression.

Our findings suggest that the retest effect is detrimental, if uncontrolled, in clinical trials for neurodegenerative diseases, such as Huntington's disease. We show here that if two assessments are performed a short time apart, use of the second assessment as the baseline increases the chances of detecting an effect of treatment, if there is one. In addition, including the retest effect in models renders the resulting models more predictive, making it possible to refine the design of future trials. This constitutes a great stride forward in cognitive assessments in clinical trials.

Supporting Information

S1 Table. Predictive model for each task.
(DOCX)

S2 Table. M matrix for calculating the 95% prediction interval for performance at A_3 for each task.
(DOCX)

S1 Text. Statistical explanation for calculation of the 95% prediction interval for performance at A_3, for each task.
(DOCX)

Acknowledgments

The authors thank Julie Sappa from Alex Edelman & Associates for her language corrections.

We thank the neurologists and the neuropsychologists from the MIGHD group trial who collected the data: A-C. Bachoud-Lévi (Henri Mondor hospital, Créteil, Principal investigator), M-F Boissé (Henri Mondor hospital, Créteil, Neuropsychologist), L. Lemoine (Henri Mondor hospital, Créteil, Neuropsychologist), C. Verny (Angers hospital, Site coordinator), G. Aubin (Angers hospital, Neuropsychologist), J-F Demonet (CHU Rangueil, Toulouse, Site coordinator), F. Calvas (CHU Rangueil, Toulouse, Investigator), P. Krystkowiak (Roger Salengro hospital, Lille and CHU d'Amiens, Amiens, Sites coordinator), C. Simonin (Roger Salengro hospital, Lille, Investigator), M. Delliaux (Roger Salengro hospital, Lille, Neuropsychologist), P. Damier (Hôpital Nord Laennec, Nantes, Site coordinator), P. Renou (Hôpital Nord Laennec, Nantes, Investigator), F. Supiot (Erasme hospital, Bruxelles, Site coordinator), H. Slama (Erasme hospital, Bruxelles, Neuropsychologist).

We thank the EHDI Study group: A-C. Bachoud-Lévi (Henri Mondor hospital Créteil, Principal investigator of the RIL ancillary study), J. S. Guillamo (Henri Mondor hospital Créteil), M-F Boissé (Henri Mondor hospital Créteil, Neuropsychologist), A. Dürr (Fédération de Neurologie, Pitié-Salpêtrière hospital, Paris), F. Bloch (Fédération de Neurologie, Pitié-Salpêtrière hospital, Paris), O. Messouak (Fédération de Neurologie, Pitié-Salpêtrière hospital, Paris), C. Tallaksen (Fédération de Neurologie, Pitié-Salpêtrière hospital, Paris), B. Dubois (Fédération de Neurologie, Pitié-Salpêtrière hospital, Paris), A. Engles (Hôpital Roger Salengro, Lille), P. Krystkowiak, (Hôpital Roger Salengro, Lille) A. Destee (Hôpital Roger Salengro, Lille), A. Memin (Hôpital Roger Salengro, Lille), S. Thibaut-Tanchou (Hôpital Roger Salengro, Lille), F. Pasquier (Hôpital Roger Salengro, Lille, Neurology), J-P. Azulay (CHU Purpan, Toulouse), M. Galitzky (CHU Purpan, Toulouse), O. Rascol (CHU Purpan, Toulouse), H. Mollion (Pierre Wertheimer hospital, Lyon), E. Broussolle (Pierre Wertheimer hospital, Lyon), M. Madigand (La Beauchée hospital, Saint-Brieuc), F. Lallement (La Beauchée hospital, Saint-Brieuc), C. Goizet (Haut-Lévêque hospital, Pessac), F. Tison (Haut-Lévêque hospital, Pessac) S. Arguillère (CHG du Pays d'Aix, Aixen-Provence), F. Viallet (CHG du Pays d'Aix, Aixen-Provence) S. Bakchine (Maison Blanche hospital, Reims), J. Khoris, (Gui de Chauliac hospital, Montpellier), M. Pages (Gui de Chauliac hospital, Montpellier), W. Camu (Gui de Chauliac hospital, Montpellier), F. Resch (Charles Nicolle hospital, Rouen), D. Hannequin (Charles Nicolle hospital, Rouen), F. Durif (Gabriel Montpied hospital, Clermont-Ferrand), D. Saudeau (CHRU Bretonneau,Tours), A. Autret (CHRU Bretonneau,Tours).

Author Contributions

Conceived and designed the experiments: ACBL CS SK. Analyzed the data: CS SK. Wrote the paper: ACBL CS SK. Collected the data: ACBL KY JFD PK FS CV LCL.

References

1. Bates G, Tabrizi S, Jones L. Huntington's Disease. 3rd ed. Oxford: Oxford University Press; 2014.
2. Bachoud-Lévi A-C, Maison P, Bartolomeo P, Boissé M-F, Dalla-Barba G, Ergis A-M, et al. Retest effects and cognitive decline in longitudinal follow-up of patients with early HD. Neurology. 2001; 56 (8):1052–8. PMID: 11320178
3. Salthouse TA, Schroeder DH, Ferrer E. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. Dev Psychol. 2004; 40(5):813–22. PMID: 15355168
4. Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc. 2003; 9(3):419–28. PMID: 12666766
5. Stout JC, Queller S, Baker KN, Cowlishaw S, Sampaio C, Fitzer-Attas C, et al. HD-CAB: a cognitive assessment battery for clinical trials in Huntington's disease. Mov Disord. 2014; 29(10):1281–8. doi: 10.1002/mds.25964 PMID: 25209258
6. Duff K, Beglinger LJ, Schultz SK, Moser D, McCaffrey R, Haase R, et al. Practice effects in the prediction of long-term cognitive outcome in three patient samples: a novel prognostic index. Arch Clin Neuropsychol. 2007; 22(1):15–24. PMID: 17142007
7. Duff K, Beglinger LJ, Moser DJ, Paulsen JS, Schultz SK, Arndt S. Predicting cognitive change in older adults: the relative contribution of practice effects. Arch Clin Neuropsychol. 2010; 25(2):81–8. doi: 10.1093/arclin/acp105 PMID: 20064816
8. Bachoud-Lévi A-C, Hantraye P, Peschanski M. Fetal neural grafts for Huntington's disease: a prospective view. Mov Disord. 2002; 17(3):439–44. PMID: 12112189
9. Landwehrmeyer GB, Dubois B, de Yébenes JG, Kremer B, Gaus W, Kraus P, et al. Riluzole in Huntington's disease: a 3-year, randomized controlled study. Ann Neurol. 2007; 62(3):262–72. PMID: 17702031
10. Kremer HPH, Hungtington Study Group. Unified Huntington's disease rating scale: reliability and consistency. Mov Disord. 1996; 11:136–42. PMID: 8684382

11. Mattis S. Mental status examination for organic mental syndrome in the elderly patient. In: Bellak L, Karasu TB, eds. Geriatric psychiatry: a handbook for psychiatrists and primary care physicians. New York: Grune & Stratton, 1976:p77–121.
12. Golden CJ. Stroop colour and word test. Age. 1978; 15:90.
13. Butters N, Wolfe J, Granholm E, Martone M. An assessment of verbal recall, recognition and fluency abilities in patients with Huntington's disease. Cortex. 1986; 22(1):11–32. PMID: 2940074
14. Cardebat D, Doyon B, Puel M, Goulet P, Joanette Y. Formal and semantic lexical evocation in normal subjects. Performance and dynamics of production as a function of sex, age and educational level. Acta Neurol Belg. 1990; 90(4):207–17. PMID: 2124031
15. Reitan RM. Validity of the trail making test as an indicator of organic brain damage. Percept Mot Skills. 1958; 8(3):271–6.
16. Zazzo R, Stambak M. Le test des deux barrages: Une épreuve de pointillage. Neuchatel, Switzerland: Delachaux et Niestlé; 1960.
17. Brandt J. The Hopkins verbal learning test: Development of a new memory test with six equivalent forms. Clin Neuropsychol. 1991; 5(2):125–42.
18. Rieu D, Bachoud-Lévi A-C, Laurent A, Jurion E, Dalla Barba G. Adaptation française du «Hopkins verbal learning test». Rev Neurol. 2006; 162(6):721–8.
19. Hocking RR. A Biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics. 1976; 32(1):1–49.
20. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In Selected Papers of Hirotugu Akaike. New York: Springer; 1998.
21. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979; 86 (2):420–8. PMID: 18839484
22. Snowden J, Craufurd D, Griffiths H, Thompson J, Neary D. Longitudinal evaluation of cognitive disorder in Huntington's disease. J Int Neuropsychol Soc. 2001; 7(1):33–44. PMID: 11253840
23. Tabrizi SJ, Reilmann R, Roos RAC, Durr A, Leavitt B, Owen G, et al. Potential endpoints for clinical trials in premanifest and early Huntington's disease in the TRACK-HD study: analysis of 24 month observational data. Lancet Neurol. 2012; 11(1):42–53. doi: 10.1016/S1474-4422(11)70263-0 PMID: 22137354
24. Benedict RH, Zgaljardic DJ. Practice effects during repeated administrations of memory tests with and without alternate forms. J Clin Exp Neuropsychol. 1998; 20(3):339–52. PMID: 9845161
25. Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken D, Crawford J, et al. Practice effects and the use of alternate forms in serial neuropsychological testing. Arch Clin Neuropsychol. 2005; 20 (4):517–29. PMID: 15896564
26. McCaffrey RJ, Westervelt HJ. Issues associated with repeated neuropsychological assessments. Neuropsychol Rev. 1995; 5(3):203–21. PMID: 8653109
27. Cooper DB, Lacritz LH, Weiner MF, Rosenberg RN, Cullum CM. Category fluency in mild cognitive impairment: reduced effect of practice in test-retest conditions. Alzheimer Dis Assoc Disord. 2004; 18 (3):120–2. PMID: 15494616
28. Cooper DB, Epker M, Lacritz L, Weiner M, Rosenberg RN, Honig L, et al. Effects of practice on category fluency in Alzheimer's disease. Clin Neuropsychol. 2001; 15(1):125–8. PMID: 11778573
29. Rosenblatt A, Kumar BV, Mo A, Welsh CS, Margolis RL, Ross CA. Age, CAG repeat length, and clinical progression in Huntington's disease. Mov Disord. 2012; 27(2):272–6. doi: 10.1002/mds.24024 PMID: 22173986
30. Stine OC, Pleasant N, Franz ML, Abbott MH, Folstein SE, Ross CA. Correlation between the onset age of Huntington's disease and length of the trinucleotide repeat in IT-15. Hum Mol Genet. 1993; 2 (10):1547–9. PMID: 8268907
31. Lee JM, Ramos EM, Lee JH, Gillis T, Mysore JS, Hayden MR, et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology. 2012; 78(10):690–5. doi: 10.1212/WNL.0b013e318249f683 PMID: 22323755
32. Langbehn DR, Hayden MR, Paulsen JS and the PREDICT-HD Investigators of the HuntingtonStudy Group. CAG-repeat length and the age of onset in Huntington disease (HD): a review and validation study of statistical approaches. Am J Med Genet B Neuropsychiatr Genet. 2010; 153B(2):397–408. doi: 10.1002/ajmg.b.30992 PMID: 19548255
33. Næs T, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. J Chemom. 2001; 15(4):413–26.
34. Marder K, Zhao H, Myers RH, Cudkowicz M, Kayson E, Kieburtz K, et al. Rate of functional decline in Huntington's disease. Neurology. 2000; 54(2):452. PMID: 10668713

35. Brandt J, Bylsma FW, Gross R, Stine OC, Ranen N, Ross CA. Trinucleotide repeat length and clinical progression in Huntington's disease. Neurology. 1996; 46(2):527–31. PMID: 8614526

36. Ruocco HH, Bonilha L, Li LM, Lopes-Cendes I, Cendes F. Longitudinal analysis of regional grey matter loss in Huntington disease: effects of the length of the expanded CAG repeat. J Neurol Neurosurg Psychiatry. 2008; 79(2):130–5. PMID: 17615168

37. Kieburtz K, MacDonald M, Shih C, Feigin A, Steinberg K, Bordwell K, et al. Trinucleotide repeat length and progression of illness in Huntington's disease. J Med Genet. 1994; 31(11):872–4. PMID: 7853373

38. Feigin A, Kieburtz K, Bordwell K, Como P, Steinberg K, Sotack J, et al. Functional decline in Huntington's disease. Mov Disord. 1995; 10(2):211–4. PMID: 7753064

39. Carcaillon L, Berrut G, Sellalm F, Dartigues J-F, Gillette S, Péré J-J, et al. Diagnosis of Alzheimer's disease patients with rapid cognitive decline in clinical practice: interest of the Deco questionnaire. J Nutr Health Aging. 2011; 15(5):361–6. PMID: 21528162

40. Noda A, Kraemer HC, Taylor JL, Schneider B, Ashford JW, Yesavage JA. Strategies to reduce site differences in multisite studies: a case study of Alzheimer disease progression. Am J Geriatr Psychiatry. 2006; 14(11):931–8. PMID: 17068315

S1 Table. Predictive model for each task

Task	Model
Letter Fluency 1’	Women: $10.27 + 0.66 \times$ score at $A_1 + 0.84 \times$ retest Men: $10.27 + 0.66 \times$ score at $A_1 + 0.84 \times$ retest $- 2.55$
Categorical Fluency 1’	Women: $6.55 + 0.57 \times$ score at $A_1 + 0.55 \times$ retest Men: $6.55 + 0.57 \times$ score at $A_1 + 0.55 \times$ retest $- 1.82$
SDMT	Women: $-0.84 + 0.98 \times$ score at $A_1 + 0.33 \times$ retest Men: $-0.84 + 0.98 \times$ score at $A_1 + 0.33 \times$ retest $- 1.93$
Stroop W	$1.56 + 0.93 \times$ score at $A_1 + 1.04 \times$ retest
Stroop C	Women: $3.03 + 1.01 \times$ score at $A_1 + 0.43 \times$ retest Men: $3.03 + 1.01 \times$ score at $A_1 + 0.43 \times$ retest $- 8.43$
Stroop C/W	Women: $2.07 + 0.97 \times$ score at $A_1 + 0.65 \times$ retest Men: $2.07 + 0.97 \times$ score at $A_1 + 0.65 \times$ retest $- 3.65$
HVLT: Immediate recall	Women and motor first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $+ 0.26 \times$ education level Women and cognitive first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $+ 0.26 \times$ education level $- 2.08$ Women and psychiatric first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $+ 0.26 \times$ education level $+ 0.45$ Men and motor first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $- 2.01 + 0.26 \times$ education level Men and cognitive first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $- 2.01 + 0.26 \times$ education level $- 2.08$ Men and psychiatric first symptom: $6.08 + 0.53 \times$ score at $A_1 + 0.27 \times$ retest $- 2.01 + 0.26 \times$ education level $+ 0.45$
HVLT: delayed recall	$-0.51 + 0.55 \times$ score at $A_1 + 0.23 \times$ education level
HVLT recognition	$-1.35 + 0.87 \times$ score at $A_1 + 0.52 \times$ retest $+ 0.19 \times$ education level
MDRS	Women, motor first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 0.39 \times$ time since onset Women, motor first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 1.12 - 0.39 \times$ time since onset Women, cognitive first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 0.39 \times$ time since onset $+ 0.18$ Women, cognitive first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 1.12 - 0.39 \times$ time since onset $+ 0.18$ Women, psychiatric first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 0.39 \times$ time since onset $+ 1.97$ Women, psychiatric first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 1.12 - 0.39 \times$ time since onset $+ 1.97$ Men, motor first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 0.39 \times$ time since onset Men, motor first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 1.12 - 0.39 \times$ time since onset Men, cognitive first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 0.39 \times$ time since onset $+ 0.18$

	Men, cognitive first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 1.12 - 0.39 \times$ time since onset $+ 0.18$ Men, psychiatric first symptom and maternal inheritance: $20.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 0.39 \times$ time since onset $+ 1.97$ Men, psychiatric first symptom and paternal inheritance: $0.29 + 0.89 \times$ score at $A_1 + 0.64 \times$ retest $- 0.10 \times$ age at $A_1 - 2.81 - 1.12 - 0.39 \times$ time since onset $+ 1.97$
1-figure cancellation	$1.51 + 0.89 \times$ score at $A_1 + 0.57 \times$ retest
2-figure cancellation	$0.24 + 0.93 \times$ score at $A_1 + 0.50 \times$ retest
3-figure cancellation	Maternal inheritance: $7.28 + 0.83 \times$ score at $A_1 + 0.55 \times$ retest $+ 0.11 \times$ age of parent at onset Paternal inheritance: $7.28 + 0.83 \times$ score at $A_1 + 0.55 \times$ retest $- 1.96 + 0.11 \times$ age of parent at onset
TMT A time	$13.11 + 0.90 \times$ score at $A_1 + 0.59 \times$ retest
TMT B time	$28.68 + 0.94 \times$ score at $A_1 + 0.86 \times$ retest $- 0.98 \times$ age at $A_1 + 0.63 \times$ age of parent at onset
Behavior	Women: $40.26 + 0.31 \times$ score at $A_1 + 0.52 \times$ retest $- 0.92 \times$ education level $- 0.68 \times$ CAG $+ 1.36 \times$ time since onset Men: $40.26 + 0.31 \times$ score at $A_1 + 0.52 \times$ retest $+ 2.91 - 0.92 \times$ education level $- 0.68 \times$ CAG $+ 1.36 \times$ time since onset
Motor	Maternal inheritance: $-32.91 + 0.81 \times$ score at $A_1 + 0.68 \times$ retest $+ 0.52 \times$ age at $A_1 - 0.23 \times$ age of parent at onset $+ 0.63 \times$ CAG Paternal inheritance: $-32.91 + 0.81 \times$ score at $A_1 + 0.68 \times$ retest $+ 0.52 \times$ age at $A_1 + 4.40 - 0.23 \times$ age of parent at onset $+ 0.63 \times$ CAG
FAS	$0.42 + 0.65 \times$ score at $A_1 + 0.03 \times$ age at $A_1 + 0.18 \times$ CAG
IS	$70.41 + 0.63 \times$ score at $A_1 - 0.22 \times$ age at $A_1 - 0.66 \times$ CAG
TFC	Maternal inheritance: $-0.55 + 0.98 \times$ score at $A_1 + 1.50 \times$ retest $- 0.003 \times$ age at A_1 Paternal inheritance: $-0.55 + 0.98 \times$ score at $A_1 + 1.50 \times$ retest $- 0.003 \times$ age at $A_1 + 0.35$

SDMT: Symbol Digit Modalities Test; Stroop C, W and C/W: Stroop color, word and color/word interference; HVLT: Hopkins Verbal Learning Task; MDRS: Mattis Dementia Rating Scale; TMT A, B: Trail-Making Test A and B; FAS: Functional Assessment Scale; IS: Independence Scale; TFC: Total Functional Capacity.

S2 Table. M matrix for calculating the 95% prediction interval for performance at A_3 for each task

l \ k	$\hat{\sigma}$	Intercept	Score at A_1	Retest (ΔA_2-A_1)	Age at A_1	Sex	Education level	Inheritance	Age of parent at onset	CAG	Time since onset	First symptom	
												Cognitive	Psychiatric
Letter fluency 1'	7.786												
Intercept		0.25358											
Score at A_1		-0.00754	0.00031										
Retest (ΔA_2-A_1)		-0.00214	-0.00001	0.00054									
Sex		-0.04983	-0.00074	0.00123		0.10401							
Categorical fluency 1'	2.915												
Intercept		0.45934											
Score at A_1		-0.02805	0.00199										
Retest (ΔA_2-A_1)		-0.01490	0.00094	0.00256									
Sex		-0.07094	0.00059	0.00048		0.10115							
SDMT	3.835												
Intercept		0.32575											
Score at A_1		-0.00969	0.00037										
Retest (ΔA_2-A_1)		-0.00438	0.00003	0.00161									

Sex		-0.06106	-0.00018	0.00138	0.10083
Stroop C	11.84				
Intercept		0.60485			
Score at A_1		-0.00796	0.00012		
Retest (ΔA_2-A_1)		-0.00453	0.00006	0.00034	
Sex		-0.08288	0.00030	0.00005	0.10035
Stroop W	8.146				
Intercept		0.51270			
Score at A_1		-0.00985	0.00040		
Retest (ΔA_2-A_1)		-0.00465	0.00008	0.00048	
Stroop C/W	5.372				
Intercept		0.3831			
Score at A_1		-0.01078	0.00037		
Retest (ΔA_2-A_1)		-0.00458	0.00010	0.00071	
Age at A_1					
Sex		-0.08905	0.00085	0.00091	0.10216
HVLT: immediate recall	3.183				
Intercept		0.93121			
Score at A_1		-0.02234	0.00108		

	Retest (ΔA_2-A_1)		-0.01596	0.00048	0.00139						
	Sex		-0.13675	0.00340	0.00047		0.12118				
	Education level		-0.02796	-0.00018	0.00045		-0.00074	0.00271			
First	Cognitive		-0.00787	0.00011	-0.00262		0.01535	-0.00428		0.18405	
symptom	Psychiatric		0.03215	-0.00163	-0.00271		-0.01555	-0.00299		0.05352	0.14796
HVLT: delayed recall		1.783									
	Intercept		0.43801								
	Score at A_1		-0.01399	0.00297							
	Education level		-0.02571	-0.00055				0.00240			
HVLT: recognition		1.144									
	Intercept		1.94602								
	Score at A_1		-0.15123	0.01467							
	Retest (ΔA_2-A_1)		-0.07514	0.00614	0.01207						
	Education level		-0.02551	-0.00037	0.00066			0.00238			
MDRS		3.463									
	Intercept		14.67372								
	Score at A_1		-0.10020	0.00074							
	Retest (ΔA_2-A_1)		-0.06547	0.00044	0.00169						
	Age at A_1		-0.01846	0.000004	0.00006	0.00040					

	Sex		-0.31087	0.00213	0.00351	-0.00122	0.12977					
	Inheritance		-0.14897	0.000002	0.00121	0.00183	-0.01876	0.12425				
	Time since onset		-0.05450	0.00021	0.00045	0.00005	0.00300	-0.00021	0.00502			
First	Cognitive		-0.09605	0.00067	-0.00613	-0.00043	-0.00234	-0.01223	-0.00303	0.22221		
symptom	Psychiatric		-0.05031	-0.00016	-0.00145	0.00076	-0.02013	0.01822	-0.00180	0.05246	0.14541	
1-figure cancellation		2.786										
	Intercept		0.30624									
	Score at A_1		-0.01580	0.00090								
	Retest (ΔA_2-A_1)		-0.01158	0.00053	0.00249							
2-figure cancellation		2.958										
	Intercept		0.27298									
	Score at A_1		-0.01401	0.00080								
	Retest (ΔA_2-A_1)		-0.00707	0.00028	0.00310							
3-figure cancellation		3.120										
	Intercept		0.77306									
	Score at A_1		-0.01394	0.00134								
	Retest (ΔA_2-A_1)		-0.02077	0.00112	0.00413							
	Inheritance		-0.09728	0.00040	0.00459			0.14470				
	Age of parent at onset		-0.01095	-0.00015	-0.00011			-0.00008	0.00032			

TMT A time	23.42								
Intercept		0.21318							
Score at A_1		-0.00272	0.00004						
Retest (ΔA_2-A_1)		-0.00115	0.00002	0.00006					
TMT B time	22.85								
Intercept		0.96053							
Score at A_1		-0.00072	0.00001						
Retest (ΔA_2-A_1)		-0.00058	0.000002	0.00003					
Age at A_1		-0.01293	-0.00002	-0.00001	0.00059				
Age of parent at onset		-0.00681	0.00001	0.00001	-0.00024		0.00038		
Motor	7.213								
Intercept		13.46612							
Score at A_1		0.01174	0.00020						
Retest (ΔA_2-A_1)		0.01188	0.00004	0.00082					
Age at A_1		-0.09857	-0.00013	-0.00014	0.00116				
Inheritance		-0.19550	-0.00049	-0.00097	0.00305	0.13709			
Age of parent at onset		0.00156	0.00003	-0.00004	-0.00029	-0.00120	0.00039		
CAG		-0.20942	-0.00028	-0.00012	0.00139	0.00103	-0.00013	0.00358	
FAS	1.245								

Intercept		19.08516					
Score at A_1		-0.25961	0.01232				
Age at A_1		-0.08986	-0.00046		0.00095		
CAG		-0.18331	-0.00107		0.00136		0.00339
IS	6.870						
Intercept		16.94631					
Score at A_1		-0.05219	0.00054				
Age at A_1		-0.07895	-0.00003		0.00074		
Number of CAG repeats		-0.19390	0.00009		0.00109		0.00307
TFC	1.023						
Intercept		4.52855					
Score at A_1		-0.34639	0.03321				
Retest (ΔA_2-A_1)		-0.27989	0.03539	0.15364			
Age at A_1		-0.00975	-0.00085	-0.00287	0.00043		
Inheritance		-0.17175	0.00392	0.01383	0.00131	0.11296	

S1 Text. Statistical explanation for the calculation of the 95% prediction interval (95%PI) for performance at A_3, for each task

Let *P* be the predicted performance of the patient at A_3. The 95%PI is given by the following formula:

$$P \pm t_{1-\frac{\alpha}{2},df} \times \sqrt{\hat{\sigma}^2 \left(1 + \sum_k \sum_l X_k X_l M_{k,l}\right)}$$

Where $t_{1-\frac{\alpha}{2},df}$ is the student quantile of order $1 - \frac{\alpha}{2} = 0.975$ and *df* is the number of degrees of freedom, defined by *df=n-p-1*, where *n* and *p* are the number of subjects and variables in the model, respectively. It can be approximated by $t_{1-\frac{\alpha}{2},df} = 2$; $\hat{\sigma}$ is the residual variance of the predicted model; *k* and *l* are the predictive factors for each task, from the following list: *intercept*, *score at* A_1, *retest* (ΔA_2-A_1), *age at* A_1, *sex*, *education level*, *inheritance*, *age of parent at onset of disease*, *number of CAG repeats*, *time since onset* and *first symptom*; *M* is a matrix defined for each task in Supplementary Table 2. *X* is an observed characteristic of a future patient such that:

- $X_{intercept}$=1;
- If the variable *k* is quantitative, X_k=value of variable;
- X_{sex} =1 if a man, 0 if a woman;
- $X_{inheritance}$=1 if paternal inheritance, 0 if maternal inheritance;
- $X_{first\ symptom} = (X_{cognitive}, X_{psychiatric})$=(0,0) if the first symptom was motor, (1,0) if it was cognitive and (0,1) if it was psychiatric.

Discussion

Ce travail de thèse a été développé autour des essais de biothérapie dans la maladie de Huntington, une maladie neurodégénérative rare, génétique, induisant une atrophie du striatum. Plus particulièrement, nous nous sommes intéressés au cas des greffes neuronales dans le striatum des patients avec l'essai clinique MIG-HD. Les bénéfices de la greffe, mesurés par des données longitudinales, sur les précédents essais cliniques, sont hétérogènes, que ce soit entre les essais ou au sein d'un même essai clinique [18]. Cette hétérogénéité est induite par le profil des patients (progression naturelle de la maladie, troubles majeurs (moteurs, cognitifs, ...), la progression de la maladie au moment de la greffe, ...) et/ou par la procédure de greffe elle-même (technique employée, chirurgien, ...). Cette hétérogénéité, associée au faible nombre de patients inclus dans les essais, ne permet pas de conclure quant à l'efficacité des greffes. Notre travail a consisté à mieux comprendre cette hétérogénéité de l'effet du traitement pour optimiser les futurs essais de greffe dans la maladie de Huntington. Il s'est effectué en deux parties :

• La première partie consistait à trouver des sous-groupes homogènes de patients pour à réponse à la greffe. Nous y avons notamment développé une méthode de clustering pour l'effet d'un traitement dans le cadre de données longitudinales (Schramm *et al.* [121]). Nous l'avons ensuite appliquée aux données de l'essai clinique MIG-HD évaluant l'effet des greffes neuronales dans la maladie de Huntington.

• La seconde partie consistait à discuter des améliorations que l'on peut apporter aux plans expérimentaux des futurs essais cliniques, en intégrant notamment des marqueurs prédictifs de l'efficacité du traitement et/ou des marqueurs pronostiques de l'évolution de la maladie afin de réduire l'hétérogénéité des groupes de traitement. Nous y avons étudié le polymorphisme COMT comme potentiel marqueur pronostique du déclin cognitif des patients atteints de la maladie de Huntington (Schramm *et al.*, soumis). Nous avons aussi comparé la puissance de différents plans expérimentaux intégrant un marqueur prédictif. De plus, nous avons évalué l'effet d'apprentissage (retest) des tests neuropsychologiques mesurant les capacités cognitives, et montré comment une double évaluation à l'inclusion dans un essai clinique permettait de s'affranchir de cet effet retest quand le critère de jugement principal est le déclin cognitif (Schramm *et al.* [122]).

Identification des sous-groupes de patients selon la réponse à un traitement

La première étape du projet a été de proposer une nouvelle méthode de clustering pour l'effet d'un traitement dans le cadre de données longitudinales. La progression de la maladie est mesurée par un score quantitatif évoluant linéairement dans le temps. Nous nous intéressons au cas où l'initiation d'un traitement induit un changement de pente dans l'évolution de ce score. Nous utilisons donc des données longitudinales pour lesquelles nous avons une pente pré-traitement et une pente post-traitement, comme c'est le cas dans l'essai clinique MIG-HD. Les méthodes de clustering paramétriques et non paramétriques déjà développées dans le but de définir des trajectoires homogènes d'individus ne s'appliquent pas à notre question [35, 123, 34, 124]. **En effet, nous cherchons à regrouper les patients ayant des changements de pentes similaires plutôt que des pentes similaires.** Ainsi les pentes pré- et post-traitement permettent de modéliser le changement de pente et donc de définir des sous-groupes de patients, mais cela ne se traduit pas nécessairement par des pentes pré- et/ou post-traitement homogènes pour chaque sous-groupe.

Nous avons donc développé une nouvelle méthode réduisant l'information des données longitudinales en une donnée transversale sur laquelle nous pouvions utiliser les méthodes de clustering classiques. Notre méthode comprend deux étapes. Dans la première, nous utilisons un modèle mixte linéaire par morceaux (avec deux pentes) pour modéliser l'ensemble des données. Ce modèle est construit à partir de trois paramètres : (i) le score du patient à l'initiation du traitement, (ii) la pente pré-traitement et (iii) la différence de pente pré- et post-traitement. Chaque paramètre est représenté par un effet moyen de tous les patients (effet fixe) auquel s'ajoute l'effet propre à chaque patient (effet aléatoire). Les effets aléatoires sont des données quantitatives correspondant à la position de chaque patient par rapport aux effets moyens et décrivant l'hétérogénéité inter-patients. Pour chaque patient, nous obtenons donc trois données quantitatives issues des effets aléatoires, chacune correspondant à un paramètre du modèle. Dans la seconde, nous utilisons ces données quantitatives issues des effets aléatoires pour réaliser le clustering. Nous nous intéressons en particulier à l'effet aléatoire associé à la différence de pente pré- et post-traitement car ce paramètre représente une mesure de l'effet du traitement. Ces deux étapes constituent la méthode, que nous avons nommée CLEB (*Clustering for Longitudinal data with Extended Baseline*).

Afin d'évaluer notre méthode, nous avons simulé plusieurs scénarios de données et calculé le pourcentage de patients bien classés par la méthode CLEB. Cette étude de simulation montre que notre algorithme est robuste face à la variabilité intra-patient, la variabilité inter-patients en terme d'évolution ainsi que la variabilité des temps de me-

sures. Nous avons conclu que l'algorithme CLEB (Schramm *et al.* [121]) a de meilleures performances lorsque la seconde étape utilise l'algorithme de clustering paramétrique basé sur un modèle de mélange gaussien ou l'algorithme des K-moyennes avec la distance euclidienne. Nous avons testé la robustesse de notre méthode en l'appliquant sur des petits échantillons. Les résultats sont robustes y compris lorsque les groupes sont déséquilibrés avec un faible pourcentage de patients répondeurs.

L'identification de profils de réponse à un traitement est un problème récurrent en médecine qui a déjà été évoqué dans le cadre de données longitudinales. Ce fut par exemple la cas pour l'analyse des réponses à l'olanzapine et au divalproex chez des patients atteints de troubles bipolaires [125] ou encore l'analyse des réponses à un placebo [126]. La méthode utilisée consistait à réduire l'espace des données longitudinales en des données transversales pour chaque individu afin d'y appliquer l'algorithme classique des K-moyennes [33]. La réduction de l'espace se fait en modélisant les données de chaque individu (par exemple avec des régressions linéaires ou des splines). Les coefficients de régression associés sont ensuite utilisés dans l'algorithme de clustering. Plus il y a de mesures par individu, et plus les fonctions modélisant les trajectoires sont précises, augmentant le nombre de paramètres de régression qu'il est possible d'utiliser dans le clustering. L'avantage de cette méthode réside dans le fait qu'elle peut être associée à beaucoup de modèles, comme par exemple les transformations de Fourier pour modéliser des données cycliques telles que les données de météo ou des données issues d'IRM fonctionnel [127]. Cependant, bien que cette méthode soit robuste lorsqu'il y a beaucoup de données par individu et peu de variabilité, nous avons montré que notre méthode CLEB produisait de meilleures performances en particulier lorsque la variabilité intra-patient était élevée. Nous avons aussi choisi de comparer notre méthode de clustering à la méthode paramétrique basée sur le principe des classes latentes [35]. Cette méthode n'a pas été utilisée dans la recherche de profil de réponse à un traitement mais dans la recherche de profils de déclin cognitifs [128]. Elle intègre la variable catégorielle non observable « sous-groupes » (classes latentes) dans un modèle mixte et estime conjointement les paramètres du modèle et la probabilité d'appartenir à chaque classe, pour tous les individus. Nous l'avons paramétrée avec le même modèle linéaire à deux pentes que celui utilisé dans notre méthode CLEB grâce au package R `lcmm`. Notre étude de simulations a montré qu'une augmentation de l'hétérogénéité de la pente pré-traitement réduisait le pourcentage de patients bien classés par cette méthode. Nous avons aussi comparé nos résultats à la méthode non paramétrique KML [34] grâce au package R `kml`. Notre étude de simulations a montré qu'une augmentation de l'hétérogénéité de la pente pré-traitement et du score à l'initiation du traitement réduisaient le pourcentage de patients bien classés par cette méthode. Cette méthode n'a pu être appliquée que pour les scénarios où les délais inter-mesures étaient identiques pour chaque individu. Or si cela correspond au plan expérimental d'un essai clinique standard,

cela ne correspond pas, ni à la réalité de l'essai MIG-HD, pour lequel des visites ont été décalées, ni à la réalité des études observationnelles.

Les sous-groupes de patients selon la réponse à un traitement peuvent être définis en associant les données relatives à l'évolution de la maladie et les caractéristiques des patients dans l'algorithme de clustering [129]. Nous avons choisi de ne pas intégrer des données autres que celles issues de l'évolution de la maladie car nous souhaitions différencier des profils de réponse au traitement indépendamment des caractéristiques des patients d'autant plus que nous n'avions, par parti pris, aucun *a priori* sur les variables influençant la réponse au traitement.

Nous avons appliqué notre méthode sur les données de l'étude MIG-HD en utilisant le score moteur de l'UHDRS, critère principal de l'étude mesurant les performances motrices des patients. En utilisant notre méthode CLEB avec une seconde étape paramétrique (modèle de mélange gaussien), nous ne mettons pas en évidence des patients répondeurs à la greffe. Cependant utiliser les critères multivariés permettent de mettre en évidence deux groupes de patients différents quant à la gravité de la maladie au moment de l'initiation du traitement. En utilisant notre méthode CLEB avec une seconde étape non paramétrique (algorithme des K-moyennes), nous pouvons construire artificiellement des sous-groupes. Ceux-ci concordent avec les sous-groupes obtenus en comparant les pentes d'évolution des patients greffés à des patients non greffés issus de la cohorte française RHLF. Le taux de concordance est de 76% si on tient compte des cas intermédiaires et de 85% si on n'en tient pas compte. Nous montrons donc qu'avec notre méthode CLEB nous pouvons aboutir à des résultats similaires sans utiliser une autre cohorte de patients. En effet, il n'y a pas toujours des cohortes de patients disponibles pour comparer l'évolution des patients traités en essais cliniques à l'évolution des patients sans traitement.

Dans le cas spécifique de l'étude MIG-HD, nous avons choisi d'identifier les sous-groupes de patients en se basant sur l'évolution du score clinique moteur. Nous aurions pu utiliser des méthodes de clustering sur les données du métabolisme striatal obtenu par TEP, comme cela a déjà été fait en oncologie [130]. Cependant, il s'agit d'une mesure disponible trois fois au cours du temps dans l'étude MIG-HD, ce qui ne permet pas de prendre en compte toute l'évolution du patient pré- et post-traitement.

Notre méthode s'inscrit dans une approche exploratoire qui permet d'identifier un sous-groupe de patients répondant le mieux à un traitement et ainsi définir un ou plusieurs marqueur(s) prédictif(s) de l'efficacité du traitement. Cependant, augmenter le nombre de marqueurs prédictifs possibles augmente le risque de première espèce α, ce qui signifie qu'un marqueur d'efficacité du traitement peut être dû au hasard. Nous recommandons alors, comme pour toute approche exploratoire, de la faire suivre par une étude confirmatoire ou de répliquer les résultats sur une autre cohorte. Au sein même de l'algorithme

CLEB, l'utilisation de plusieurs stratégies peut aider à valider les sous-groupes quand elles convergent toutes vers le même résultat. Lorsque les différentes stratégies aboutissent à des résultats discordants, l'étude de simulations permet de générer des hypothèses sur les données et donc de choisir la meilleure stratégie. Comme nous l'avons fait dans l'article pour les données réelles, il est possible de répliquer l'algorithme sur des sous-échantillons de données afin de tester la robustesse des résultats.

Amélioration des plans expérimentaux dans les essais cliniques

La seconde étape du projet a été de proposer des pistes pour l'amélioration des futurs essais cliniques de biothérapie dans la maladie de Huntington, en y introduisant les connaissances acquises dans les études antérieures, en particulier lorsqu'on s'intéresse à un critère cognitif. Nous avons considéré trois axes.

Premièrement, nous avons montré que le polymorphisme Val^{158}Met sur le gène COMT était un marqueur pronostique du déclin cognitif chez les patients Huntington. Nous discutons ici comment ce résultat doit être confirmé et ce qu'il apporterait à la fois dans le soin courant des patients et dans la mise en place des futurs essais cliniques.

Deuxièmement, dans l'éventualité où un marqueur prédictif du traitement serait identifié, nous avons comparé les plans expérimentaux stratégiques intégrant un marqueur prédictif de l'efficacité du traitement. Nous discutons ici des précautions à prendre avant d'utiliser ces plans expérimentaux dans le cas de petits effectifs.

Troisièmement, nous avons montré que l'effet retest (amélioration des performances à la seconde évaluation) ne permettait pas de mesurer le déclin cognitif des patients en un an lorsqu'il n'y avait pas de mesure intermédiaire. Nous discutons ici de la nécessité d'utiliser une double évaluation à l'inclusion pour éviter l'effet retest.

Prise en compte des marqueurs pronostiques

Comprendre l'hétérogénéité de l'évolution de la maladie permet de mieux prendre en charge les patients, voire d'améliorer leur suivi thérapeutique. Dans le cas de la maladie de Huntington, nous avons montré que les patients ne suivaient pas tous le même déclin cognitif selon le nombre de répétitions CAG et selon leur polymorphisme Val^{158}Met sur le gène COMT. En effet, nous montrons que les patients homozygotes Met/Met ont un déclin cognitif plus rapide que les patients Val/Val. Sachant que la COMT joue un rôle dans la régulation de la dopamine, ces analyses ont permis d'émettre une hypothèse sur l'impact de la dopamine dans les capacités cognitives des patients atteints de la maladie de Huntington. De cela a émergé une proposition de prise en charge thérapeutique par neuroleptiques pour les patients Val/Val en début de maladie et par dopamine pour les patients Met/Met en fin de maladie.

Le rôle de la dopamine dans le cerveau et notamment dans la maladie de Huntington n'est pas encore clairement défini. Les connaissances actuelles semblent s'accorder sur un impact de la dopamine sur les fonctions cognitives des individus. En effet le polymorphisme Val^{158}Met du gène COMT a un impact sur l'activité de la dopamine dans le cortex préfrontal, une zone associée aux fonction cognitives supérieures (par exemple le langage, la mémoire de travail, le raisonnement, les fonctions exécutives) [103]. Ainsi, il n'est pas surprenant de voir un impact de ce polymorphisme sur le déclin cognitif des patients.

Les recherches de marqueurs pronostiques sont fréquentes mais souvent biaisées, d'où l'importance de valider les résultats que nous avons obtenus dans d'autres cohortes de patients Huntington dans des études prospectives. Ces études prospectives doivent être encadrées par des échelles d'évaluation de la qualité des études comme c'est déjà le cas en cancérologie. On peut citer en exemple les critères REMARK (*REporting recommandations for tumour MARKer prognostic studies*) pour les études tumorales [131]. La première étude à utiliser une validation prospective d'un marqueur pronostique a été réalisée dans le cadre du cancer du sein où le marqueur pronostique (score omique) visait à déterminer si la chimiothérapie était nécessaire [132]. Lorsqu'un marqueur est défini pronostique, il doit montrer une utilité clinique et/ou la faisabilité de sa mesure pour justifier son utilisation [133] en soin courant et/ou dans les essais cliniques.

Par exemple, notre hypothèse, selon laquelle le polymorphisme Val^{158}Met sur le gène COMT constitue un marqueur pronostique permettant d'ajuster le traitement, doit être validée par un essai clinique randomisé où la procédure de mesure du marqueur pronostique est spécifiée dans le protocole [96]. L'essai clinique comparera un groupe où le marqueur pronostique est utilisé pour guider le choix du traitement et un groupe où il n'est pas utilisé.

En plus d'être utilisé dans le soin courant, le marqueur pronostique peut intervenir à quatre niveaux de la conception d'un essai clinique [134, 135].

Tout d'abord, il peut être utilisé dans la sélection des patients comme critère d'inclusion et d'exclusion. Ainsi par l'augmentation de l'homogénéité des patients inclus dans l'étude, on peut diminuer la variabilité observée. Dans le cas d'un essai à petit effectif, il s'agit d'un moyen d'augmenter la puissance sans augmenter le nombre de patients à inclure.

Au moment de la randomisation, le marqueur pronostique peut devenir critère de stratification. Toujours dans notre exemple du polymorphisme Val^{158}Met sur le gène COMT dans la maladie de Huntington, cela permettrait d'homogénéiser les groupes de randomisation sur ce critère lorsqu'est étudié l'impact d'un nouveau traitement sur le déclin cognitif des patients. En effet, dans le cas de petits effectifs, la stratification permet d'éviter des déséquilibres sur les facteurs pronostiques. Les conséquences d'une stratification

seront une plus grande puissance statistique et une meilleure précision [136] ainsi qu'un contrôle du risque de première espèce α [137].

Le marqueur pronostique peut aussi être utilisé comme facteur d'ajustement. Même lorsque la randomisation a été stratifiée sur le marqueur pronostique, cet ajustement permet de mesurer l'effet du traitement avec une meilleure précision [138, 139].

Enfin le marqueur pronostique peut intervenir dans le choix du traitement, comme nous l'avons montré dans le cas du polymorphisme $Val^{158}Met$ sur le gène COMT dans la maladie de Huntington.

Prise en compte des marqueurs prédictifs

Comme pour les marqueurs pronostiques, les marqueurs prédictifs permettent d'améliorer les soins et les essais cliniques. Cependant, beaucoup de biais entourent la découverte des marqueurs prédictifs et les essais cliniques qui permettent de les valider doivent être encadrés par des lignes directrices. Nous avons fait une revue des différents plans expérimentaux basés sur un marqueur prédictif et donné leurs limites dans le cas spécifiques de petits échantillons suite à une étude de simulations.

Les marqueurs prédictifs améliorent les soins lorsqu'ils permettent de ne donner le traitement qu'aux patients qui en auront un bénéfice. Pour cela, ils doivent être validés en essais cliniques. Il faut prouver que le marqueur est un « modificateur » de l'effet du traitement. Il faut aussi montrer son utilité clinique via un plan expérimental stratégique comparant la stratégie selon laquelle le traitement est donné en fonction du marqueur et une stratégie standard [108, 109, 112]. Nous avons comparé les plans stratégiques en terme de nombre de sujets nécessaires et montré que le plan stratégique inverse [111] était le moins couteux en nombre de sujets. Nous avons quantifié l'impact de la valeur pronostique du marqueur prédictif sur la puissance des études. Nous avons montré que cela ne nuisait pas à la puissance en cas de grands effectifs. Mais, dans le cas de petits effectifs, la puissance pouvait être diminuée, de façon non monotone selon la valeur pronostique du marqueur. Enfin, nous avons montré qu'utiliser ces plans expérimentaux lorsque le marqueur n'est pas prédictif de l'effet du traitement pouvait tout de même conclure à l'utilisation du marqueur dans le choix du traitement. Nous avons quantifié la probabilité d'aboutir à la mauvaise conclusion selon la prévalence du marqueur et l'effet du traitement. Nous concluons que pour choisir le plan optimal, il est nécessaire de tenir compte à la fois des limites statistiques, des limites éthiques et bien sûr de la question posée.

Les marqueurs prédictifs permettent d'améliorer les essais cliniques, soit en incluant uniquement les patients avec un bénéfice potentiel du traitement, soit en choisissant le traitement en fonction du marqueur. Dans les deux cas, cela permettra de mettre plus facilement en évidence un effet bénéfique du traitement sur un sous-groupe pour lequel il est réellement bénéfique que sur la population générale.

Actuellement, les marqueurs prédictifs testés en essais cliniques et utilisés dans les soins courants sont essentiellement développés en cancérologie. Il s'agit le plus souvent de marqueurs génétiques présents dans les cellules tumorales, comme c'est le cas par exemple avec le cancer du sein et le marqueur BluePrint® [140] ou le cancer colorectal avec le marqueur K-RAS [141]. Mais les marqueurs prédictifs ne sont pas uniquement utilisés en cancérologie. Par exemple, le marqueur IL28B (marqueur génétique génome général) pour la réponse au traitement pegylated interferon combiné avec la ribavirin est utilisé pour guider le traitement de l'hépatite C [142]. L'utilisation de tels marqueurs permet de ne donner le traitement qu'aux patients qui en tireront un bénéfice. Ils permettent aussi de mettre en évidence un traitement efficace dans un sous-groupe de patients là où l'essai clinique sur la population plus large ne permettait pas de mettre en évidence un traitement partiellement efficace. Par exemple pour l'utilisation du gefitinib dans le cancer du poumon, différentes études ont abouti à des résultats très hétérogènes [143, 144]. Finalement des études incluant un marqueur prédictif EGFR ont montré l'efficacité du traitement pour les patients présentant la mutation [145]. Les marqueurs ne sont pas uniquement génétiques, mais peuvent être par exemple issus de l'imagerie. C'est le cas de l'électroencéphalographie (qui mesure l'activité électrique du cerveau) qui permet de discriminer les patients répondeurs ou non répondeurs à la stimulation magnétique transcrânienne pour le traitement de la dépression [146]. De même il existe une relation entre le métabolisme cérébral mesuré par TEP et l'effet des antidepresseurs chez les patients souffrants de dépression [147].

Dans le cas de l'étude MIG-HD, nous n'avons pas mis en évidence de marqueur prédictif d'un bénéfice de la greffe. Les plans stratégiques ne doivent être utilisés que lorsqu'un fort *a priori* existe sur la valeur prédictive d'un marqueur. Nous ne préconisons donc pas de se tourner vers ces plans expérimentaux pour le prochain essai greffe dans la maladie de Huntington.

Prise en compte du l'effet « retest » dans le suivi cognitif

Les troubles cognitifs sont rarement évalués en essais cliniques de par la difficulté de montrer un déclin cognitif dans un suivi longitudinal, compte tenu des meilleures performances des patients lors de la seconde passation, dues à la familiarisation des patients avec le test (effet retest) [53]. Nous nous sommes donc intéressés à mettre en évidence cet effet retest et comment il peut être éviter et/ou utiliser dans les futurs essais cliniques.

Nous avons confirmé que dans le cadre d'un suivi longitudinal, à la seconde évaluation, les patients ont de meilleures performances cognitives suite à la familiarisation avec le test (Schramm *et al.* [122]). Ce retest n'apparait pas pour les tests moteur et fonctionnels. Nous montrons qu'une double évaluation à l'inclusion dans le suivi permet de

mieux observer le déclin cognitif du patient en utilisant la seconde évaluation comme ligne de base. Nous avons montré que la mesure du retest, en plus d'autres variables, pouvait prédire les performances des patients un an plus tard. Grâce à un algorithme « pas à pas » nous avons trouvé les meilleurs prédicteurs des performances à un an. Suivre des patients en pré-traitement, pendant un an, comme c'est le cas dans l'essai MIG-HD permet de rajouter des critères d'exclusion supplémentaires et d'inclure des patients au déclin plus homogène. Nos modèles de prédiction permettent de comparer pour chaque patient, ses performances réelles à un an et ses performances théoriques calculées par le modèle. Dans le cas où le patient est un déclineur rapide, il pourra être exclu du protocole [148]. D'autre part, la définition de déclineurs rapides ou déclineurs lents grâce à ces modèles permettra aussi de stratifier la randomisation des essais cliniques sur cette caractéristique. Enfin, on pourrait imaginer que les performances théoriques à un an soient utilisées pour comparer l'effet d'un traitement où chaque patient serait son propre témoin.

Le petit échantillon ne permet pas d'estimer la performance à un an avec précision. L'intervalle de confiance associé reste large. Cependant, nous avons utilisé le groupe placebo d'un autre essai clinique (RIL-HD) dont le plan expérimental sur la première année de suivi était proche de celui de MIG-HD et nous avons montré que nos modèles étaient reproductibles sur cette seconde cohorte, ce qui a permis de les valider. Cependant, il existe un biais du fait que certains patients avaient déjà été confrontés à ces tests avant leur inclusion dans MIG-HD. De plus, au vu des résultats concernant le polymorphisme $Val^{158}Met$ sur le gène COMT, nous pouvons émettre l'hypothèse que ce marqueur pronostique améliorerait nos prédictions. Cependant, nous n'avons pas pu l'utiliser car il n'était pas déterminé lorsque l'essai MIG-HD a débuté.

L'effet retest apparaît dans la plupart des tests cognitifs des patients atteints de troubles cognitifs ou des sujets sains [149]. Cependant l'effet retest pour une même tâche peut varier selon la population évaluée [150]. Par exemple, Cooper et al [151, 152] ont montré l'existence d'un effet retest pour la tâche de fluence catégorielle chez des sujets sains mais pas chez des patients atteints de la maladie d'Alzheimer ou de troubles cognitifs légers. De même, nous n'avons trouvé aucun effet de retest pour cette tâche chez les patients atteints de la maladie de Huntington. L'effet retest pourrait être une mesure complémentaire de la performance brute du patient. En effet, l'utilisation du retest serait pertinent pour estimer une nouvelle mesure combinant à la fois le niveau de performance brute et la capacité de se familiariser avec la tâche, notamment dans les tâches nécessitant un fort niveau d'exigences cognitives [153]. L'utilisation de formes parallèles permet aussi d'amoindrir l'effet retest, notamment dans les tâches de mémoire [154]. Mais elles ne permettent pas de limiter la familiarisation avec la tâche et n'ont aucun intérêt dans les tâches avec une forte composante motrice [150].

Le retest nous empêche d'observer le déclin réel du patient. Nous suggérons d'utiliser la double évaluation à l'inclusion des patients afin d'atténuer l'effet retest et d'homogénéiser les patients à l'inclusion dans les futurs essais cliniques. Cependant, il faut continuer à étudier l'effet retest afin d'évaluer s'il s'agit d'un effet permanent ou s'atténuant avec le nombre de passations. Nous proposons donc, en perspective, de réaliser un suivi de cohorte avec double évaluation à un mois d'écart, annuellement. Cela permet de modéliser le retest et son évolution au cours du temps évolution en tenant compte du déclin cognitif et du nombre de passations déjà réalisées. Pour ce faire, il faudrait établir ce protocole à la fois chez des patients Huntington et chez des sujets sains. Outre le retest, la variabilité intrinsèque du patient est un frein à l'observation d'un déclin cognitif. En effet, les performances du patient sont impactées par son état « émotionnel ». Ainsi, le score mesuré à un instant t fluctue autour de la valeur réelle. Si le score du patient est mesuré à différents instants, la moyenne des mesures pourrait mieux refléter le score réel du patient. Pour un suivi tous les ans ou tous les six mois, on peut imaginer qu'une mesure soit réalisée plusieurs fois au cours du même mois. Même s'il paraît compliqué de faire venir le patient plusieurs fois à l'hôpital, on peut imaginer que le recueil de données multiples ne sera plus un problème quand les patients pourront être testés chez eux, notamment grâce à des outils connectés. Par une étude de simulation, nous pourrions voir l'impact de ces mesures répétées sur l'observation d'un déclin.

Conclusion générale et perspectives

Les futurs essais de biothérapie dans la maladie de Huntington

Les prochains essais de biothérapie de la maladie de Huntington comporteront une double évaluation à l'inclusion dans l'étude. L'effet retest sera mesuré et sera inclus dans des équations permettant de prédire si le patient est un déclineur rapide ou lent. Ainsi la randomisation pourra être stratifiée sur cette caractéristique ainsi que sur le polymorphisme $Val^{158}Met$ du gène COMT. Les essais greffes ne s'évalueront pas par un plan expérimental basé sur un marqueur prédictif car nous n'avons pas mis de tels marqueurs en évidence. Cependant, si de tels marqueurs sont mis en évidence pour l'effet d'une autre thérapie, les essais pourront choisir le plan expérimental le plus approprié sur la base de notre étude de simulations. De plus, dans le cas de thérapie génique, le protocole devrait préciser la possibilité de récupérer des informations génétiques relatives aux patients afin d'identifier de potentiels marqueurs prédictifs.

Extension de la méthode CLEB au cas multivarié pour définir de nouveaux sous-groupes dans l'étude MIG-HD

Nous n'avons pas pu mettre en évidence de marqueur prédictif de l'effet des greffes dans la maladie de Huntington. Les sous-groupes de patients ont été définis à partir de l'évolution de leurs performances au test moteur de l'UHDRS. Cependant, la maladie de Huntington comprend des symptômes variés et peut-être que le score moteur ne peut pas refléter à lui seul l'efficacité de la greffe. Nous souhaitons poursuivre l'analyse des données de l'essai MIG-HD en incluant les performances des patients aussi bien dans le domaine moteur, que dans les domaines cognitifs, psychiatriques et fonctionnels. Pour cela, nous souhaitons étendre notre méthode CLEB au cas multivarié. Deux perspectives s'offrent à nous. Premièrement, nous pouvons réaliser un modèle mixte à deux pentes sur chacun des p tests d'intérêt et appliquer les méthodes de clustering sur la matrice des effets aléatoires correspondant au changement de pente des p modèles pour chaque patient. Une seconde possibilité serait d'utiliser un modèle multivarié à classes latentes où le modèle mixte à deux pentes ne modéliserait plus un score observé mais cette variable latente non observable modélisant la progression de la maladie, les scores observés étant quant à eux des transformations linéaires ou non linéaires de cette variable latente [106]. Ainsi, les effets aléatoires utilisés dans le clustering correspondraient à l'évolution de la variable latente. Une méthode multivariée pourrait peut-être mettre en évidence de nouveaux sous-groupes de réponse au traitement des greffes grâce aux données de MIG-HD.

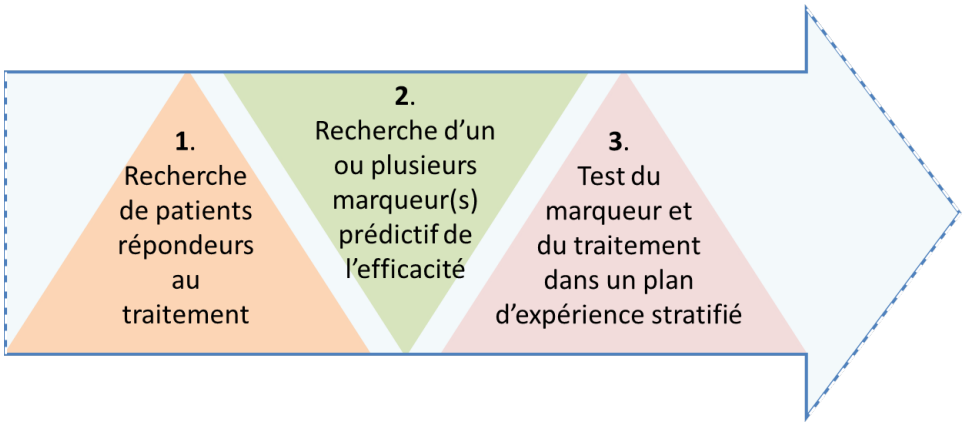


FIGURE 34 – Les étapes de la médecine stratifiée

1. La première étape consiste à trouver les sous-groupes de patients répondeurs au traitement, notamment grâce à l'algorithme CLEB. 2. La deuxième étape consiste à définir des marqueurs prédictifs de l'efficacité du traitement. 3. La troisième étape consiste à valider ces marqueurs et les traitements pour la population pouvant être des éventuels répondeurs.

CLEB : un pas vers la médecine stratifiée

Depuis toujours la médecine cherche à améliorer la prise en charge des patients en étant de plus en plus précise dans les conditions d'administration des traitements (choix de la molécule, choix de la dose, durée du traitement, ...). Elle se dirige vers une médecine personnalisée, qui vise à donner le meilleur traitement en fonction des caractéristiques biologiques et génétiques des individus. Cependant, la mise en place d'un traitement individualisé reste complexe et utopique [155]. La médecine stratifiée, trop souvent confondue avec la médecine personnalisée, tente d'identifier des sous-groupes de patients homogènes par rapport à la réponse à un traitement donné [156]. Elle s'appuie sur des marqueurs pronostiques de l'évolution de la maladie et des marqueurs prédictifs de l'efficacité du traitement.

L'algorithme CLEB permet de définir des sous-groupes de patients selon l'impact d'un traitement, ce qui constitue un pas vers la médecine stratifiée [157, 156]. En effet, après avoir défini des sous-groupes de patients répondeurs, nous pouvons chercher les caractéristiques des patients qui expliquent la réponse au traitement (voir Figure 34). Le plus souvent, les recherches se tournent vers des marqueurs génétiques. Ces marqueurs peuvent ensuite être utilisés dans de futurs essais cliniques afin de valider leur valeur prédictive et l'efficacité du traitement dans le groupe possédant ce marqueur génétique. Ces nouveaux plans expérimentaux ont particulièrement émergés dans le domaine de la cancérologie. Ils permettent de réduire le nombre de sujets nécessaires, ce qui représente un atout pour les maladies rares.

Annexes

Annexe A

Echelles d'évaluation et plans expérimentaux utilisés dans la maladie de Huntington

A.1 Les échelles d'évaluation de la maladie de Huntington

L'échelle de référence internationale pour le suivi des patients Huntington est l'UHDRS (*Unified Huntington's Disease Rating Scale*) [13]. Elle est composée de sous-échelles motrice, fonctionnelles, psychiatrique et cognitives.

Les capacités motrices sont évaluées par :

- le test moteur (*Total Motor Score*, TMS) évoluant de 0 à 124 où 0 correspond à aucun trouble moteur et 124 à un trouble moteur maximal.

Les capacités fonctionnelles sont évaluées par :

- le test de capacité fonctionnelle (*Total Functional Capacity*, TFC) évoluant de 0 à 13 où 13 correspond à aucun trouble fonctionnel et 0 à un trouble fonctionnel maximal ;
- le test d'appréciation fonctionnelle (*Functional Assessment Scale*, FAS) évoluant de 25 à 50 où 25 correspond à aucun trouble fonctionnel et 50 à un trouble fonctionnel maximal ;
- le test de dépendance (*Independance Scale*, IS) évoluant de 100 à 0 où 100 correspond à aucun trouble fonctionnel et 0 à un trouble fonctionnel maximal.

Les troubles psychiatriques sont évalués par :

- le test psychiatrique évoluant de 0 à 88 où 0 correspond à aucun trouble psychiatrique et 88 à un trouble psychiatrique maximal.

Les capacités cognitives sont évaluées par :

- le test de fluence littérale mesurant le nombre de mots donnés en une ou deux

minutes débutant par les lettres P, R et V ;

- le test des symboles (*Symbol Digit Modalities Test*, SDMT) consistant à remplacer des chiffres par des symboles géométriques qui leur sont attribués. Les performances sont mesurées en nombre de réponses correctes données en 90 secondes ;
- le test de Stroop, divisé en trois partie. La première (test des couleurs) consiste à dénommer des couleurs à partir de rectangles de couleurs, la seconde (test des mots) à lire des noms de couleurs écrits en noir et la troisième (test d'interférence) à dénommer les couleurs dans laquelle sont écrit des noms de couleur différent de la couleur à dénommer. Les performances sont mesurées nombre de réponses correctes données en 90 secondes pour chaque partie [158].

En plus des tests de l'UHDRS, les patients atteints de la maladie de Huntington de la cohorte RHLF ou de l'essai MIG-HD sont optionnellement évalués par une échelle de dépression :

- le test de de Montgomery et Asberg (*Montgomery and Asberg Depression Rating Scale*, MADRS) évoluant de 0 à 60 où 0 correspond à aucun trouble dépressif et 60 à un trouble dépressif maximal.

ainsi que des tests cognitifs :

- le test de Mattis (*Mattis Dementia Rating Scale*, MDRS) évoluant de 0 à 144 où 144 correspond à aucun trouble cognitif et 0 à un trouble cognitif maximal [159]. Cette échelle est composée de cinq sous-échelles testant l'attention, la persévération, la capacité de construction, de conceptualisation, et la mémoire ;
- le test de fluence catégorielle mesurant le nombre de mots appartenant à une même catégorie (exemple : animaux) produits en une ou deux minutes [160, 161] ;
- le test de barrage de signes de Zazzo mesurant le nombre de signes corrects barrés en 90 secondes pour un, deux ou trois signes à barrer [162] ;
- le test de mémoire de Hopkins (*Hopkins Verbal Learning Test*, HLVT) qui consiste à apprendre une liste de 12 mots devant être retenus et restitués en rappel immédiat, en rappel différé ou en test de reconnaissance [163, 164].
- le *Trail Making Test* (TMT), versions A et B, comptant le nombre de points reliés correctement en 240 secondes, chacun sur 25 points. Dans la version A il s'agit de relier une suite de nombres (1-2-3-...) et dans la version B il s'agit de relier une suite alternant nombres et lettres (1-A-2−-B-...) [165].

A.2 Les essais cliniques dans la maladie de Huntington

Nous avons réalisé une requête sur clinicalTrial.gov en utilisant les termes « Huntington » ou « Chorea » (autre nom associé à la maladie de Huntington) pour déterminer les caractéristiques des études enregistrées comme incluant des patients atteints de la maladie de Huntington. Cette requête a aboutit à 130 études dont 108 incluaient réellement des patients atteints de la maladie de Huntington et non des sujets sains ou des patients atteints d'autres troubles induisant des chorées. Sur ces 108 études, il y a 77 essais cliniques et 31 études observationnelles. Nous nous sommes intéressés aux plans expérimentaux et aux critères de jugement utilisés dans ces 77 essais cliniques.

Table 9 – Description des plans expérimentaux utilisés dans les essais cliniques portant sur la maladie de Huntington

	Phase							
	0 N=1	I N=8	I/II N=6	II N=33	II/III N=3	III N=12	IV N=3	non renseigné N=11
Plans randomisés								
Plan parallèle	–	3	4	17	3	9	–	2
Plan « cross-over »	–	2	–	3	–	–	–	–
Plan « delayed-start »	–	–	–	5	–	1	–	–
Plans non randomisés								
Plan avec un seul traitement	1	3	2	6	–	2	3	7
Plan « cross-over »	–	–	–	2	–	–	–	2

Le plan randomisé en parallèle est le plan expérimental le plus utilisé en essais cliniques. Les plans avec un seul traitement sont pour 7 d'entre eux des poursuites d'études dont le but est d'évaluer l'effet toxique à long terme du traitement proposé au bras de randomisation « traitement expérimental » d'un essai précédent.

TABLE 10 – Description des critères de jugement utilisés dans les essais cliniques portant sur la maladie de Huntington

	Phase								Total
	0	I	I/II	II	II/III	III	IV	non renseigné	
	N=1	N=8	N=6	N=33	N=3	N=12	N=3	N=11	
Critère principal									
Moteur	–	1	–	8	1	5	–	–	15
Fonctionnel	–	–	–	–	1	3	–	–	4
Cognitif	–	–	1	3	–	1	2	1	8
Biologique	–	3	1	–	–	–	–	2	4
Cérébral	1	2	1	3	–	1	–	–	8
Toxicité	–	–	4	16	–	2	–	1	23
Autres	–	2	1	7	1	1	2	7	21
Critère secondaire									
Moteur	–	1	1	14	–	3	–	3	22
Fonctionnel	–	–	–	–	–	2	–	–	2
Cognitif	–	1	1	12	1	3	–	1	19
Biologique	–	1	1	9	–	–	–	–	11
Cérébral	1	1	–	4	–	–	–	–	6
Toxicité	–	–	1	3	–	2	–	–	6
Autres	–	–	1	15	–	6	1	4	27

Les essais de phase III ont pour but d'évaluer l'effet du traitement sur les symptômes de la maladie de Huntington. Les critères principaux sont essentiellement moteur ou fonctionnel puisqu'ils définissent ce que le patient est capable de réaliser malgré ses mouvements anormaux. Les critères cognitifs, tout aussi importants sont préférés en critères de jugements secondaires.

Annexe B

Clustering pour l'effet d'un traitement sur des événements récurrents

Un événement récurrent est un événement pouvant survenir plusieurs fois chez un même individu au cours du temps [166]. Les événements successifs peuvent être identiques (exemple : crise d'épilepsie) ou de gravité ordonnée (exemples : entrée dans un nouveau stade d'une maladie, détérioration de l'acuité visuelle). Ces données sont dites censurées car on ne connaît pas, de manière exhaustive, tous les événements d'un patient. Par exemple, la censure à droite intervient avec la fin de la collecte des données de l'étude ou parce que survient un événement absorbant (exemple : décès du patient). Les modèles de durée, aussi appelés modèles de survie permettent d'analyser ce type de données. Nous avons adapté notre méthode CLEB au cas des événements récurrents afin de construire des sous-groupes de patients selon leur réponse à un traitement, la réponse étant évaluée par l'occurrence des événements. Cette nouvelle méthode CREME (*Clustering for Recurrent Event using Mixed Effects*) est aussi constituée de deux étapes. La première étape consiste à modéliser le délai de survenu des événements en fonction du traitement où le traitement est une variable binaire dépendante du temps, grâce à une adaptation du modèle de Cox pour les événements récurrents intégrant des effets aléatoires. La seconde étape consiste à utiliser les estimation des effets aléatoire du modèle comme entrées dans un algorithme de clustering classique pour données transversales.

Dans un premier temps nous décrivons la modélisation des événements récurrents, en particulier lors de l'évaluation d'un effet du traitement dépendant du temps. Dans un second temps nous décrirons la méthode CREME. Enfin, nous montrerons les résultats de notre étude de simulation évaluant les performances de notre méthode.

B.1 Modélisation des événements récurrents

Hypothèse

Nous nous plaçons dans le cas où les événements sont identiques et indépendants, c'est-à-dire que la probabilité de faire un événement est identique quelque soit le nombre d'événements déjà rencontrés par l'individu.

Notations

Soit $i \in \{1,...,N\}$ un individu observé jusqu'au temps C_i (temps de censure). On note T_{ik} le temps d'apparition réel du $k$ème événement de l'individu i. On observe le couple de variables (Y_{ik},δ_{ik}) où $Y_{ik} = min(T_{ik},C_i)$ et $\delta_{ik} = 1$ si $T_{ik} \leq C_i$ et $\delta_{ik} = 0$ sinon (variable de censure). Par défaut, $Y_{i0} = 0$. Soit $G_{ik} = Y_{ik} - Y_{i,k-1}$ le délai entre les observations $k-1$ et k. Enfin, $\lambda_0(t)$ représente le risque de base de survenu d'un événement au temps t. Notons que sous notre hypothèse, le risque de base ne dépend pas du nombre d'événements déjà rencontrés par l'individu.

Modélisation

Le modèle de Cox, utilisé pour les données de survie classiques [167], a été étendu pour l'analyse des événements récurrents. Kelly et Lim [168] ont comptabilisé ainsi sept modèles possibles dont trois correspondant à notre hypothèse : le modèle de Andersen et Gill (AG) [169], le modèle « *Gap Time - UnRestricted* » (GT-UR) et le modèle de Lee, Wei et Amato (LWA) [170]. La différence entre ces trois modèles est la définition de l'intervalle de risque. Le modèle AG considère un processus de comptage, ainsi les intervalles de risque sont définis à partir des temps observés (Figure 35.A). Le modèle GT-UR considère le délai entre les événements en débutant chaque nouvel intervalle de risque à 0 (Figure 35.B). Le modèle LWA considère le délai total entre l'entrée dans l'étude et l'apparition de l'événement (Figure 35.C). La table 11 résume les temps de début et de fin des intervalles de risque pour les données fictives présentées sur la figure 35.

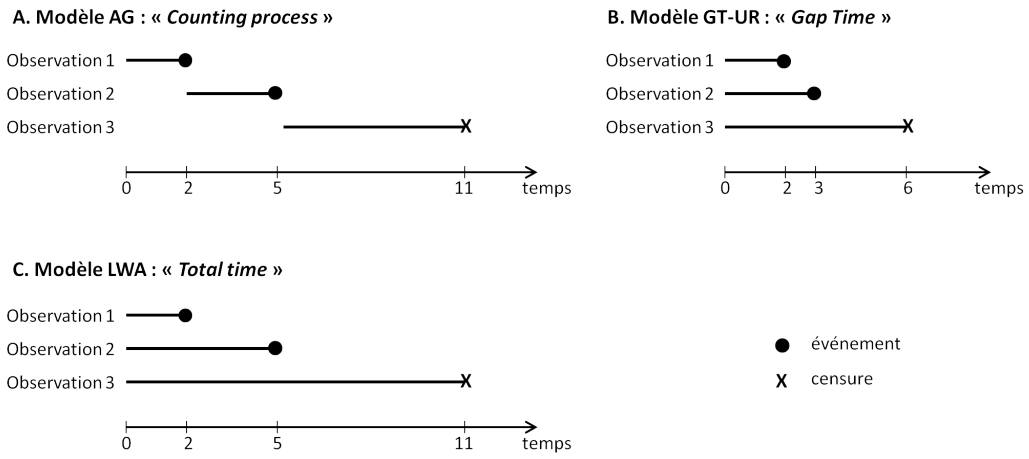


FIGURE 35 – Définition des intervalles de risque

Ces données sont issues des trois observations d'un même individu. Celui-ci a rencontré deux événements aux temps 2 et 5 et il a été suivi jusqu'au temps 11 qui représente une censure.

TABLE 11 – Début et fin des intervalles de risque selon le modèle

	AG		GT-UR		LWA		
	Début	Fin	Début	Fin	Début	Fin	δ
Observation 1	0	2	0	2	0	2	1
Observation 2	2	5	0	3	0	5	1
Observation 3	5	11	0	6	0	11	0

Ces données sont issues des trois observations d'un même individu, présentées par la figure 35. Celui-ci a rencontré deux événements aux temps 2 et 5 et il a été suivi jusqu'au temps 11 qui représente une censure.

Ainsi, au temps t, l'individu i est considéré à risque pour son $k$ème événement si $Y_{i,k-1} < t \leq Y_{i,k}$ (modèle AG) ou si $t < G_{i,k}$ (modèle GT-UR) ou si $t < Y_{i,k}$ (modèle LWA).

Soit X la matrice des covariables et $\lambda_{ik}(t,X_i)$ le risque qu'a l'individu i de présenter le $k$ème événement au temps t sachant ses caractéristiques X_i. Notons β le vecteur des coefficients de régression associés à X. Les équations B.1 et B.2 modélisent respectivement le risque de survenu d'un événement par les modèles AG ou LWA et GT-UR.

$$\text{Modèle AG ou LWA :} \qquad \lambda_{ik}(t,X_i) = \lambda_0(t)e^{\beta X_i} \tag{B.1}$$

$$\text{Modèle GT-UR :} \qquad \lambda_{ik}(t,X_i) = \lambda_0(t - t_{k-1})e^{\beta X_i} \tag{B.2}$$

Mesurer l'effet d'un traitement sur l'occurrence des événements

Soit τ_i la date d'initiation du traitement pour l'individu i. Alors la covariable informant de la prise du traitement s'écrit sous forme d'une indicatrice : $X_i(t) = \mathbb{1}(t \geq \tau_i)$ qui varie en fonction du temps. Pour tenir compte de cette fonction du temps, les intervalles de risque doivent être redécoupés en ajoutant une observation à chaque fois que la variable change de valeur, comme proposé dans le tableau 12.

TABLE 12 – Début et Fin des intervalles de risque selon le modèle quand l'individu initie un traitement au temps 2,5

	AG		GT-UR		LWA			
	Début	Fin	Début	Fin	Début	Fin	$X_i(t)$	δ
Observation 1	0	2	0	2	0	2	0	1
Traitement	2	2,5	0	0,5	0	2,5	0	0
Observation 2	2,5	5	0	2,5	0	5	1	1
Observation 3	5	11	0	6	0	11	1	0

Ces données sont issues des trois observations d'un même individu, présentées par la figure 35. Celui-ci a rencontré deux événements aux temps 2 et 5 et il a été suivi jusqu'au temps 11 qui représente une censure. Il a eu le traitement au temps 2,5.

Modèles à fragilité

Les modèles présentés ci-dessus ne tiennent pas compte de la corrélation intra-patient, sauf si un estimateur robuste de la variance est utilisé (estimateur « sandwich ») [171]. Mais en cas de forte corrélation, l'effet du traitement peut-être sous-estimé. Il convient alors d'utiliser plutôt des modèles de fragilité consistant à ajouter un effet aléatoire proportionnel au risque de base [172]. Dans ce cas, les modèles AG, LWA et GT-UR sont modifiés de la façon suivante :

Modèle AG ou LWA :

$$\lambda_{ik}(t,X_i) = \lambda_0(t)\eta e^{\beta X_i} \;\Leftrightarrow\; \lambda_{ik}(t,X_i) = \lambda_0(t)e^{\upsilon+\beta X_i} \text{ où } \eta = e^{\upsilon} \tag{B.3}$$

Modèle GT-UR :

$$\lambda_{ik}(t,X_i) = \lambda_0(t-t_{k-1})\eta e^{\beta X_i} \;\Leftrightarrow\; \lambda_{ik}(t,X_i) = \lambda_0(t-t_{k-1})e^{\upsilon+\beta X_i} \text{ où } \eta = e^{\upsilon} \tag{B.4}$$

Le terme de fragilité η suppose un risque de base variable entre les individus [173]. La loi gamma est l'hypothèse la plus standard pour la distribution de η. Mais d'autres lois peuvent être considérées comme la loi inverse gaussienne ou encore la loi positive stable. Les hypothèses sous-jacentes sont une plus forte corrélation entre les événements tardifs (loi gamma) ou au contraire entre les événements précoces (loi positive stable). L'écriture $\upsilon + \beta X_i$ correspond à un terme mixte où β est un effet fixe et υ un effet aléatoire. A

la différence des modèles mixtes pour données longitudinales, les effets aléatoires ne sont pas considérés comme ayant une distribution normale. Cependant, par analogie, avec les modèles mixtes pour données longitudinales, des modèles faisant l'hypothèse $\upsilon_i \sim \mathcal{N}(0,\sigma_\upsilon^2)$ ont été développés [174]. Dans ces modèles à effets mixtes les effets aléatoires sont supposés suivre une loi gaussienne et peuvent être associés à l'intercept, ce qui revient à l'ajout d'un terme de fragilité, ou associés à des covariables du modèle.

B.2 La méthode CREME (*Clustering for Recurrent Event using Mixed Effects*)

Problématique

L'effet du traitement peut être hétérogène avec par exemple seulement un sous-groupe de patients pouvant en retirer un bénéfice. Si l'effet du traitement est mesuré par son impact sur l'apparition des événements, le sous-groupe de patients ayant un bénéfice du traitement verra l'occurrence des événements diminuer voire ne fera plus d'événement tandis que les autres continueront à en faire autant voire d'avantage. Trouver ces sous-groupes de patients est important pour mieux définir les patients à exposer au traitement et comprendre pourquoi le traitement n'est pas efficace chez certains afin d'améliorer leur prise en charge. Cette problématique est similaire à celle qui a abouti à construire la méthode CLEB.

La première étape de notre méthode consiste à modéliser les données en intégrant la variable traitement et des effets aléatoires associés au risque de base et au traitement comme le montre les équations B.5 et B.6.

Modèle AG et LWA :

$$\lambda_{ik}(t,\tau_i) = \lambda_0(t)e^{\upsilon_i+(\beta+\omega_i)\mathbb{1}(t\geq\tau_i)} \tag{B.5}$$

Modèle GT-UR :

$$\lambda_{ik}(t,\tau_i) = \lambda_0(t-t_{k-1})e^{\upsilon_i+(\beta+\omega_i)\mathbb{1}(t\geq\tau_i)} \tag{B.6}$$

où β correspond au coefficient associé à l'effet du traitement, τ_i est la date de l'initiation du traitement pour le patient i, λ_0 es le risque de base, υ est l'effet aléatoire associé au risque de base tel que $\upsilon_i \sim \mathcal{N}(0,\sigma_\upsilon^2)$ et ω est l'effet aléatoire associé à l'effet du traitement tel que $\omega_i \sim \mathcal{N}(0,\sigma_\omega^2)$. Le tableau 13 résume le parallèle entre la modélisation des données longitudinales continues et la modélisation des données d'événements récurrents.

TABLE 13 – Parallèle entre les données longitudinales continues et les événements récurrents pour l'extension de la méthode CLEB

	Données longitudinales continues	Evénements récurrents
Données d'intérêt	pente d'évolution	occurrence des événements
Effet positif du traitement	diminution de la pente	diminution du nombre d'événements et augmentation des délais d'apparition
Variabilité inter-individus avant traitement	pente pré-traitement + score à l'initiation du traitement	risque de base
	$\rightarrow$ effets aléatoires b_0 et b_1	$\rightarrow$ effet aléatoire υ
Hétérogénéité de l'effet du traitement	différents impacts du traitement sur le changement de pente	différents impacts du traitement sur l'occurrence des événements
	$\rightarrow$ effet aléatoire b_2	$\rightarrow$ effet aléatoire w

La seconde étape de notre méthode consiste à utiliser les estimations de ω_i en entrée d'un méthode de clustering classique pour données transversales.

B.3 Etude de simulation

Nous avons mis en place une étude de simulation afin d'évaluer les différentes stratégies possibles au sein de la méthode CREME en combinant différents modèles à l'étape 1 et différentes méthode de clustering à l'étape 2.

Génération des données pour l'étude de simulation

Les données pour l'étude de simulation ont été générées à partir d'un échantillon de 200 patients. Nous supposons deux sous-groupes de patients. Le groupe A de taille N_A a un effet bénéfique du traitement. Le groupe B de taille N_B n'a pas d'effet du traitement.

On note τ_i la date d'initiation du traitement pour le patient i, qui a été générée par une loi uniforme : $\tau_i \sim \mathcal{U}(\tau_{\min},\tau_{\max})$. L'effet du traitement est supposé constant dans le

temps. On suppose une censure aléatoire à droite modélisée par $C_i \sim \mathcal{U}(300,500)$.

Pour chaque patient i, notons T_{ik} le temps réel d'apparition du $k$ème événement de l'individu i, où $T_{i0} = 0$, ce qui signifie que le patient entre dans l'étude lorsqu'il fait son premier événement. Soit Y_{ik} les observations de chaque patient. Alors $Y_{ik} = \min(T_{ik},C_i)$. Le délai entre chaque observation d'événement a été généré en utilisant une fonction exponentielle tenant compte de la variable traitement dont la valeur varie au cours du suivi [175].

Pour chaque patient i les observations sont générées de la façon suivante :

- $T_{i0} = 0$ et $Y_{i0} = 0$
- $k = 1$
- Tant que $T_{i,k-1} < C_i$
  - $u \sim \mathcal{U}(0,1)$
  - si $T_{i,k-1} < \tau_i$, alors :

$$T_{i,k}-T_{i,k-1} = \begin{cases} \frac{-ln(u)}{\lambda_0 exp(\beta_{0i}^g)} & \text{si} -ln(u) < \lambda_0 exp(\beta_{0i}^g)(\tau_i - T_{i,k-1}) \\ \frac{-ln(u)-\lambda_0 exp(\beta_{0i}^g)(\tau_i-T_{i,k-1})+\lambda exp(\beta_{0i}^g+\beta_{1i}^g)(\tau_i-T_{i,k-1})}{\lambda_0 exp(\beta_{0i}^g+\beta_{1i}^g)} & \text{si} -ln(u) \geq \lambda exp(\beta_{0i}^g)(\tau_i - T_{i,k-1}) \end{cases}$$

  - si $T_{i,k-1} \geq \tau_i$: $T_{i,k} - T_{i,k-1} = \frac{-ln(u)}{\lambda exp(\beta_{0i}^g+\beta_{1i}^g)}$
  - $Y_{ik} = \min(T_{ik},C_i)$
  - $k \leftarrow k + 1$

Les données ont été simulées avec les valeurs des paramètres suivantes : $(\beta_0^{(g)},\beta_1^{(g)}) \sim \mathcal{N}(\mu^{(g)},\Sigma^{(g)})$ avec $\mu^{(1)} = (ln(0.01),ln(0.01))$, $\mu^{(2)} = (ln(0.01),0)$, $\Sigma^{(1)} = \Sigma^{(2)} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}$, $\tau_{\min} = 190$ et $\tau_{\max} = 210$.

Résultats

Le résultats de l'étude de simulation sont présentés sur la figure 36 en terme de pourcentage de patients correctement classés. Dans chaque scénario l'utilisation du modèle LWA ne donne pas de bon résultats. Les modèles AG et GT-UR donnent des résultats similaires mais le modèle AG semble avoir de meilleures performances en particulier lorsque les sous-groupes sont déséquilibrés (Figure 36.E). Le nombre total de patients n'affecte pas la méthode qui est donc robuste dans le cas de faibles effectifs (Figure 36.D). Lorsque la différence d'effet du traitement entre les deux sous-groupes diminue, le pourcentage de patients correctement classés diminue jusqu'à atteindre un taux de 50% lorsqu'il n'y a pas de différence entre les deux sous-groupes, ce qui correspond à un classement « au hasard » (Figure 36.A). On observe aussi une diminution du pourcentage de patients correctement classés lorsque la variabilité liée à l'effet du traitement augmente.

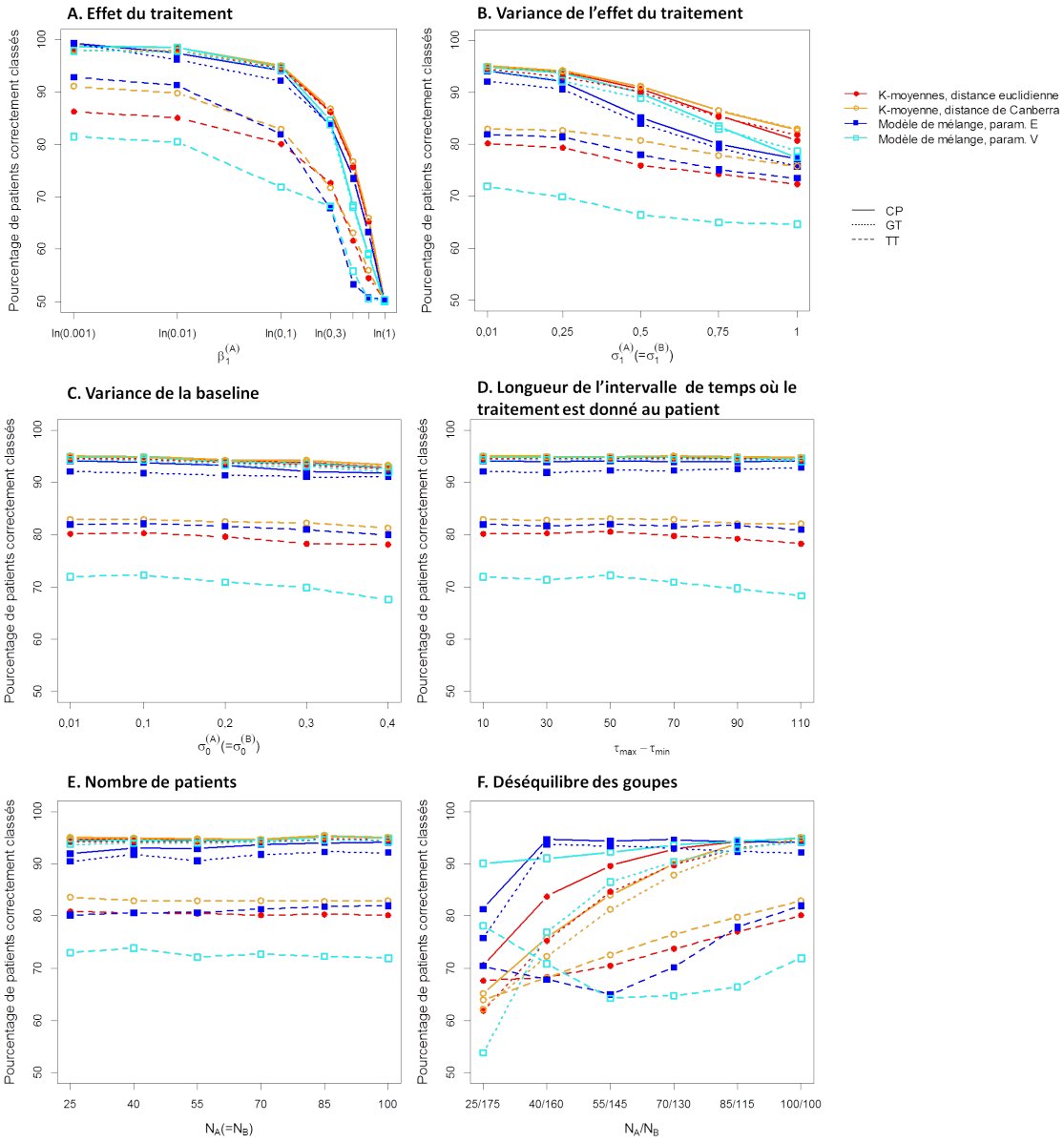


FIGURE 36 – Pourcentage de patients correctement classés avec la méthode CREME

Discussion

L'étude de simulation montre des résultats similaires entre notre méthode CREME et notre méthode CLEB. Nos résultats montrent que nous ne pouvons pas utiliser le modèle LWA au sein de la méthode CREME. Ce résultat est en accord avec l'étude de simulation de Kelly et Lim [168] montrant que ce modèle n'était pas approprié pour l'étude des événements récurrents car il conduit à une estimation biaisée de l'effet du traitement. Ce modèle peut s'étendre au cas où l'effet du traitement n'est pas immédiat. En effet s'il

existe un délai τ' entre l'initiation du traitement et son effet, il peut être pris en compte en remplaçant le dans le modèle, τ_i par $\tau_i + \tau'$. Nous avons supposé une distribution normale des effets aléatoires. Nous souhaitons poursuivre notre étude de simulation dans le cas où la distribution des effets aléatoires n'est pas générée par une loi normale. De même, nous avons généré la censure selon une loi uniforme. Nous souhaiterons évaluer l'impact d'une censure générée par une loi exponentielle sur les performances de notre méthode. De plus, nous avons supposé des censures à droites qui sont les censures le plus souvent rencontrées dans les bases de données réelles. Cependant, nous aimerions savoir comment notre méthode peut être adaptée en présence de censure à gauche.

Annexe C

Calcul de la puissance comme une fonction de la valeur pronostique du marqueur

C.1 Notations et puissance du test

Notations

Soit π, la prévalence du marqueur $M+$ et n, le nombre de patients dans chaque bras de stratégie. Nous nous plaçons dans le cas où le traitement standard n'a aucun effet (indépendamment du marqueur) et où le traitement expérimental n'a aucun effet chez les patients $M-$. Nous avons :

- θ_0 : probabilité de faire l'événement pour un patient $M-$ ne recevant aucun traitement
- θ_{M+} : effet additionnel du marqueur $M+$, indépendamment du traitement (effet pronostique)
- $\theta_{Exp} = 0$: effet additionnel du traitement expérimental, indépendamment du marqueur
- $\theta_{Sdt} = 0$: effet additionnel du traitement standard, indépendamment du marqueur
- θ_{Exp+} : effet d'interaction entre le traitement expérimental et le marqueur $M+$ (effet prédictif)
- $\theta_{Sdt+} = 0$: effet d'interaction entre le traitement standard et le marqueur $M+$ (effet prédictif)

Alors les probabilités de faire l'événement selon le marqueur et le traitements sont :

- $p_{Exp+} = \theta_0 + \theta_{M+} + \theta_{Exp} + \theta_{Exp+} = \theta_0 + \theta_{M+} + \theta_{Exp+}$: probabilité de faire l'événement pour un patient $M+$ recevant le traitement expérimental
- $p_{Exp-} = \theta_0 + \theta_{Exp} = \theta_0$: probabilité de faire l'événement pour un patient $M-$ recevant le traitement expérimental

- $p_{Std+} = \theta_0 + \theta_{M+} + \theta_{Std} + \theta_{Std+} = \theta_0 + \theta_{M+}$: probabilité de faire l'événement pour un patient $M+$ recevant le traitement standard
- $p_{Std-} = \theta_0 + \theta_{Std} = \theta_0$: probabilité de faire l'événement pour un patient $M-$ recevant le traitement standard

La probabilité de faire l'événement pour un patient inclus dans le bras stratégique (respectivement dans le bras contrôle) est notée p_S (respectivement p_C.

Puissance du test

Nous décrivons ici comment écrire la puissance du test :

$$1 - \beta = \mathbb{P}\left(u \ \leq \ \frac{\sqrt{n/2}\,(p_S - p_C)}{\sqrt{p_S(1-p_S) + p_C(1-p_C)}} - Z_{\alpha/2} \right) , \qquad u \sim \mathcal{N}(0,1) \tag{C.1}$$

sous la forme :

$$1 - \beta = \mathbb{P}\left(u \ \leq \ \frac{\sqrt{n}\gamma}{\sqrt{a + b\theta_{M+} + c\theta_{M+}^2}} - Z_{\alpha/2} \right) , \qquad u \sim \mathcal{N}(0,1) \tag{C.2}$$

C.2 Cas des plans expérimentaux stratégiques simple et inverse

Pour les plans stratégiques simple et inverse, nous avons, dans le bras stratégique : $p_S = \pi p_{Exp+} + (1-\pi)p_{Std-} = \theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+}$. Dans le bras contrôle, nous avons $p_C = \pi p_{Std+} + (1-\pi)p_{Std-} = \theta_0 + \pi\theta_{M+}$ pour le plan stratégique simple, et $p_C = \pi p_{Std+} + (1-\pi)p_{Exp-} = \theta_0 + \pi\theta_{M+}$ pour le plan stratégique inverse.

Alors :

$$\begin{aligned} p_S - p_C &= \theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+} - \theta_0 - \pi\theta_{M+} \\ &= \pi\theta_{Exp+} \end{aligned}$$

$$\begin{aligned} p_S(1-p_S) + p_C(1-p_C) &= (\theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+})(1 - \theta_0 - \pi\theta_{M+} - \pi\theta_{Exp+}) \\ &\quad + (\theta_0 + \pi\theta_{M+})(1 - \theta_0 - \pi\theta_{M+}) \\ &= 2\theta_0 - 2\theta_0^2 + \pi\theta_{Exp+} - 2\pi\theta_0\theta_{Exp+} - \pi^2\theta_{Exp+}^2 + (2\pi - 4\pi\theta_0 - 2\pi^2\theta_{Exp+})\theta_{M+} - 2\pi^2\theta_{M+}^2 \end{aligned}$$

Ce qui revient aux valeurs de paramètres suivantes :

$$\begin{cases} \gamma & = \frac{\pi\theta_{Exp+}}{\sqrt{2}} \\ a & = 2\theta_0 - 2\theta_0^2 + \pi\theta_{Exp+} - 2\pi\theta_0\theta_{Exp+} - \pi^2\theta_{Exp+}^2 \\ b & = 2\pi - 4\pi\theta_0 - 2\pi^2\theta_{Exp+} \\ c & = -2\pi^2 \end{cases} \tag{C.3}$$

C.3 Cas du plan expérimental stratégique modifié

Pour le plan stratégique modifié, nous avons, dans le bras stratégique : $p_S = \pi p_{Exp+} + (1-\pi)p_{Std-} = \theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+}$, et dans le bras contrôle : $p_C = \frac{\pi}{2}(p_{Exp+} + p_{Std+}) + \frac{1-\pi}{2}(p_{Exp-} + p_{Std-}) = \theta_0 + \pi\theta_{M+} + \frac{1}{2}\pi\theta_{Exp+}$.

Alors :

$$\begin{aligned} p_S - p_C &= \theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+} - \theta_0 - \pi\theta_{M+} - \tfrac{1}{2}\pi\theta_{Exp+} \\ &= \tfrac{1}{2}\pi\theta_{Exp+} \end{aligned}$$

$$\begin{aligned} p_S(1-p_S) + p_C(1-p_C) &= (\theta_0 + \pi\theta_{M+} + \pi\theta_{Exp+})(1 - \theta_0 - \pi\theta_{M+} - \pi\theta_{Exp+}) \\ &\quad + (\theta_0 + \pi\theta_{M+} + \tfrac{1}{2}\pi\theta_{Exp+})(1 - \theta_0 - \pi\theta_{M+} - \tfrac{1}{2}\pi\theta_{Exp+}) \\ &= 2\theta_0 - 2\theta_0^2 + \pi\theta_{Exp+} - 3\pi\theta_0\theta_{Exp+} - \tfrac{5}{4}\pi^2\theta_{Exp+}^2 + (2\pi - 4\pi\theta_0 - 2\pi^2\theta_{Exp+} + \tfrac{1}{2}\pi\theta_{Exp+} \\ &\quad - \pi^2\theta_{Exp+})\theta_{M+} - 2\pi^2\theta_{M+}^2 \end{aligned}$$

Ce qui revient aux valeurs de paramètres suivantes :

$$\begin{cases} \gamma & = \frac{\pi\theta_{Exp+}}{2^{3/2}} \\ a & = 2\theta_0 - 2\theta_0^2 + \pi\theta_{Exp+} - 3\pi\theta_0\theta_{Exp+} - \frac{5}{4}\pi^2\theta_{Exp+}^2 \\ b & = 2\pi - 4\pi\theta_0 + (\frac{1}{2} - 3\pi)\pi\theta_{Exp+} \\ c & = -2\pi^2 \end{cases} \tag{C.4}$$

Bibliographie

[1] Shoulson I and Fahn S. Huntington disease clinical care and evaluation. *Neurology*. 1979. 29(1) : 1–1.

[2] MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*. 1993. 72(6) : 971–983.

[3] Squitieri F. Neurodegenerative disease : 'fifty shades of grey' in the huntington disease gene. *Nature Reviews Neurology*. 2013. 9(8) : 421–422.

[4] Nance MA, Seltzer W, Ashizawa T, Bennett R, McIntosh N, Myers RH, *et al.* Laboratory guidelines for huntington disease genetic testing. *The American Journal of Human Genetics*. 1998. 62(5) : 1243–1247.

[5] Squitieri F and Jankovic J. Huntington's disease : how intermediate are intermediate repeat lengths ? *Movement Disorders*. 2012. 27(14) : 1714–1717.

[6] Stine OC, Pleasant N, Franz ML, Abbott MH, Folstein SE and Ross CA. Correlation between the onset age of huntington's disease and length of the trinucleotide repeat in IT-15. *Human molecular genetics*. 1993. 2(10) : 1547–1549.

[7] Langbehn DR, Hayden MR and Paulsen JS. CAG-repeat length and the age of onset in huntington disease (hd) : a review and validation study of statistical approaches. *American Journal of Medical Genetics Part B : Neuropsychiatric Genetics*. 2010. 153(2) : 397–408.

[8] Telenius H, Kremer H, Thellmann J, Andrew S, Almqvist E, Anvret M, *et al.* Molecular analysis of juvenile huntington disease : the major influence on (CAG) n repeat length is the sex of the affected parent. *Human molecular genetics*. 1993. 2(10) : 1535–1540.

[9] Ribaï P, Nguyen K, Hahn-Barma V, Gourfinkel-An I, Vidailhet M, Legout A, *et al.* Psychiatric and cognitive difficulties as indicators of juvenile huntington disease onset in 29 patients. *Archives of neurology*. 2007. 64(6) : 813–819.

[10] Broadstock M, Michie S and Marteau T. Psychological consequences of predictive genetic testing : a systematic review. *European journal of human genetics : EJHG*. 2000. 8(10) : 731–738.

[11] Gauthier LR, Charrin BC, Borrell-Pagès M, Dompierre JP, Rangone H, Cordelières FP, *et al.* Huntingtin controls neurotrophic support and survival of neurons by enhancing bdnf vesicular transport along microtubules. *Cell*. 2004. 118(1) : 127–138.

[12] Pringsheim T, Wiltshire K, Day L, Dykeman J, Steeves T and Jette N. The incidence and prevalence of huntington's disease : A systematic review and meta-analysis. *Movement Disorders*. 2012. 27(9) : 1083–1091.

[13] Kremer H, Group HS, *et al.* Unified huntington's disease rating scale : reliability and consistency. 1996.

[14] Schoenfeld M, Myers RH, Cupples LA, Berkman B, Sax DS and Clark E. Increased rate of suicide among patients with huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*. 1984. 47(12) : 1283–1287.

[15] Verma IM and Somia N. Gene therapy-promises, problems and prospects. *Nature*. 1997. 389(6648) : 239–242.

[16] Hoffman D, Gitlitz BJ, Belldegrun A and Figlin RA. Adoptive cellular therapy. In Seminars in oncology, vol. 27. 2000 pp. 221–233.

[17] Fraser JK, Schreiber RE, Zuk PA and Hedrick MH. Adult stem cell therapy for the heart. *The international journal of biochemistry & cell biology*. 2004. 36(4) : 658–666.

[18] Bachoud-Lévi AC and Perrier A. Regenerative medicine in huntington's disease : Current status on fetal grafts and prospects for the use of pluripotent stem cell. *Revue neurologique*. 2014. 170(12) : 749–762.

[19] Galpern WR, Corrigan-Curay J, Lang AE, Kahn J, Tagle D, Barker RA, *et al.* Sham neurosurgical procedures in clinical trials for neurodegenerative diseases : scientific and ethical considerations. *The Lancet Neurology*. 2012. 11(7) : 643–650.

[20] Dekkers W and Boer G. Sham neurosurgery in patients with parkinson's disease : is it morally acceptable ? *Journal of Medical Ethics*. 2001. 27(3) : 151–156.

[21] Lillie EO, Patay B, Diamant J, Issell B, Topol EJ and Schork NJ. The n-of-1 clinical trial : the ultimate strategy for individualizing medicine ? *Personalized medicine*. 2011. 8(2) : 161–173.

[22] Palfi S, Condé F, Riche D, Brouillet E, Dautry C, Mittoux V, *et al.* Fetal striatal allografts reverse cognitive deficits in a primate model of huntington disease. *Nature medicine*. 1998. 4(8) : 963–966.

[23] Dunnett SB, Carter R, Watts C, Torres EM, Mahal A, Mangiarini L, *et al.* Striatal transplantation in a transgenic mouse model of huntington's disease. *Experimental neurology*. 1998. 154(1) : 31–40.

[24] Kopyov O, Jacques S, Lieberman A, Duma C and Eagle K. Safety of intrastriatal neurotransplantation for huntington's disease patients. *Experimental neurology*. 1998. 149(1) : 97–108.

[25] Bachoud-Lévi AC, Bourdet C, Brugieres P, Nguyen JP, Grandmougin T, Haddad B, *et al.* Safety and tolerability assessment of intrastriatal neural allografts in five patients with huntington's disease. *Experimental neurology*. 2000. 161(1) : 194–202.

[26] Barker RA, Mason SL, Harrower TP, Swain RA, Ho AK, Sahakian BJ, *et al.* The long-term safety and efficacy of bilateral transplantation of human fetal striatal tissue in patients with mild to moderate huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*. 2013. 84(6) : 657–665.

[27] Reuter I, Tai YF, Pavese N, Chaudhuri KR, Mason S, Polkey CE, *et al.* Long-term clinical and positron emission tomography outcome of fetal striatal transplantation in huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*. 2008. 79(8) : 948–951.

[28] Gallina P, Paganini M, Di Rita A, Lombardini L, Moretti M, Vannelli GB, *et al.* Human fetal striatal transplantation in huntington's disease : a refinement of the stereotactic procedure. *Stereotactic and functional neurosurgery*. 2008. 86(5) : 308–313.

[29] Paganini M, Biggeri A, Romoli AM, Mechi C, Ghelli E, Berti V, *et al.* Fetal striatal grafting slows motor and cognitive decline of huntington's disease. *Journal of Neurology, Neurosurgery & Psychiatry*. 2013. pp. jnnp–2013.

[30] Hauser RA, Furtado S, Cimino C, Delgado H, Eichler S, Schwartz S, *et al.* Bilateral human fetal striatal transplantation in huntington's disease. *Neurology*. 2002. 58(5) : 687–695.

[31] Bachoud-Lévi AC, Gaura V, Brugières P, Lefaucheur JP, Boissé MF, Maison P, *et al.* Effect of fetal neural transplants in patients with huntington's disease 6 years after surgery : a long-term follow-up study. *The Lancet Neurology*. 2006. 5(4) : 303–309.

[32] Everitt BS, Landau S, Leese M and Stahl D. Cluster Analysis, U.K. : Wiley-Blackwell : Chichester, West Sussex2011, 5th ed.

[33] Tarpey T and Kinateder KK. Clustering functional data. *Journal of classification*. 2003. 20(1) : 093–114.

[34] Genolini C and Falissard B. Kml : k-means for longitudinal data. *Computational Statistics*. 2010. 25(2) : 317–328.

[35] Muthén B and Muthén LK. Integrating person-centered and variable-centered analyses : Growth mixture modeling with latent trajectory classes. *Alcoholism : Clinical and experimental research*. 2000. 24(6) : 882–891.

[36] Koestler DC, Marsit CJ, Christensen BC, Kelsey KT and Houseman EA. A recursively partitioned mixture model for clustering time-course gene expression data. *Translational cancer research*. 2014. 3(3) : 217.

[37] Harrington M, Velicer WF and Ramsey S. Typology of alcohol users based on longitudinal patterns of drinking. *Addictive behaviors*. 2014. 39(3) : 607–621.

[38] Castellini G, Fioravanti G, Sauro CL, Rotella F, Lelli L, Ventura L, *et al.* Latent profile and latent transition analyses of eating disorder phenotypes in a clinical sample : A 6-year follow-up study. *Psychiatry research*. 2013. 207(1) : 92–99.

[39] Kent P and Kongsted A. Identifying clinical course patterns in sms data using cluster analysis. *Chiropractic & manual therapies*. 2012. 20(1) : 1–12.

[40] Tepper PG, Randolph Jr JF, McConnell DS, Crawford SL, El Khoudary SR, Joffe H, *et al.* Trajectory clustering of estradiol and follicle-stimulating hormone during the menopausal transition among women in the study of women's health across the nation (swan). *The Journal of Clinical Endocrinology & Metabolism*. 2012. 97(8) : 2872–2880.

[41] Cornu C, Kassai B, Fisch R, Chiron C, Alberti C, Guerrini R, *et al.* Experimental designs for small randomised clinical trials : an algorithm for choice. *Orphanet J Rare Dis*. 2013. 8(1) : 48.

[42] Guyatt GH, Heyting A, Jaeschke R, Keller J, Adachi JD and Roberts RS. N of 1 randomized trials for investigating new drugs. *Controlled clinical trials*. 1990. 11(2) : 88–100.

[43] Scuffham PA, Nikles J, Mitchell GK, Yelland MJ, Vine N, Poulos CJ, *et al.* Using n-of-1 trials to improve patient management and save costs. *Journal of general internal medicine*. 2010. 25(9) : 906–913.

[44] Zucker D, Schmid C, McIntosh M, D'Agostino R, Selker H and Lau J. Combining single patient (n-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of clinical epidemiology*. 1997. 50(4) : 401–410.

[45] Stallard N and Rosenberger WF. Exact group-sequential designs for clinical trials with randomized play-the-winner allocation. *Statistics in medicine*. 2002. 21(4) : 467–480.

[46] Zhang L, Chan WS, Cheung SH and Hu F. A generalized drop-the-loser urn for clinical trials with delayed responses. *Statistica Sinica*. 2007. 17(1) : 387.

[47] Sun R, Cheung SH and Zhang LX. A generalized drop-the-loser rule for multi-treatment clinical trials. *Journal of Statistical Planning and Inference*. 2007. 137(6) : 2011–2023.

[48] Bhattaram VA, Siddiqui O, Kapcala LP and Gobburu JV. Endpoints and analyses to discern disease-modifying drug effects in early parkinson's disease. *The AAPS journal*. 2009. 11(3) : 456–464.

[49] Clarke CE. Are delayed-start design trials to show neuroprotection in parkinson's disease fundamentally flawed ? *Movement Disorders*. 2008. 23(6) : 784–789.

[50] Olanow CW, Rascol O, Hauser R, Feigin PD, Jankovic J, Lang A, *et al.* A double-blind, delayed-start trial of rasagiline in parkinson's disease. *New England Journal of Medicine*. 2009. 361(13) : 1268–1278.

[51] Carcaillon L, Berrutp G, Sellalm F, Dartigues JF, Gillette S, Péré JJ, *et al.* Diagnosis of alzheimer's disease patients with rapid cognitive decline in clinical practice : interest of the deco questionnaire. *The journal of nutrition, health & aging*. 2011. 15(5) : 361–366.

[52] Kang M, Ragan BG and Park JH. Issues in outcomes research : an overview of randomization techniques for clinical trials. *Journal of Athletic Training*. 2008. 43(2) : 215.

[53] Bachoud-Lévi AC, Maison P, Bartolomeo P, Boissé MF, Dalla Barba G, Ergis AM, *et al.* Retest effects and cognitive decline in longitudinal follow-up of patients with early hd. *Neurology*. 2001. 56(8) : 1052–1058.

[54] Siesling S, van Vugt JP, Zwinderman KA, Kieburtz K and Roos RA. Unified huntington's disease rating scale : a follow up. *Movement disorders*. 1998. 13(6) : 915–919.

[55] Salmon DP, Thal LJ, Butters N and Heindel WC. Longitudinal evaluation of dementia of the alzheimer type a comparison of 3 standardized mental status examinations. *Neurology*. 1990. 40(8) : 1225–1225.

[56] Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*. 1977. pp. 1–38.

[57] Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982. pp. 963–974.

[58] Madsen K, Miller J and Province M. The use of an extended baseline period in the evaluation of treatment in a longitudinal duchenne muscular dystrophy trial. *Statistics in medicine*. 1986. 5(3) : 231–241.

[59] Fomin SV, *et al.* Elements of the theory of functions and functional analysis, vol. 1, Courier Corporation1999.

[60] Cha SH. Comprehensive survey on distance/similarity measures between probability density functions. *City*. 2007. 1(2) : 1.

[61] Larson RC and Sadiq G. Facility locations with the manhattan metric in the presence of barriers to travel. *Operations Research*. 1983. 31(4) : 652–669.

[62] Lance GN and Williams WT. Mixed-data classificatory programs i - agglomerative systems. *Australian Computer Journal*. 1967. 1(1) : 15–20.

[63] Rokach L and Maimon O. Chapter 15—clustering methods. *The Data Mining and Knowledge Discovery Handbook*. 2013. pp. 321–352.

[64] Hartigan JA and Wong MA. Algorithm as 136 : A k-means clustering algorithm. *Applied statistics*. 1979. 28(1) : 100–108.

[65] MacQueen J, *et al.* Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, Oakland, CA, USA.1967 pp. 281–297.

[66] Selim SZ and Ismail MA. K-means-type algorithms : a generalized convergence theorem and characterization of local optimality. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 1984. (1) : 81–87.

[67] Arthur D and Vassilvitskii S. k-means++ : The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics2007 pp. 1027–1035.

[68] Bahmani B, Moseley B, Vattani A, Kumar R and Vassilvitskii S. Scalable k-means++. *Proceedings of the VLDB Endowment*. 2012. 5(7) : 622–633.

[69] Kaufman L and Rousseeuw P. Statistical Data Analysis Based on the Ll Norm, chap. Clustering by means of medoids, pp. 405–416, Amsterdam : North-Holland1987, y. dodge ed.

[70] Estivill-Castro V and Yang J. Fast and robust general purpose clustering algorithms. In PRICAI 2000 Topics in Artificial Intelligence, pp. 208–218, Springer2000.

[71] Hans-Hermann B. Origins and extensions of the k-means algorithm in cluster analysis. *Journal Electronique d'Histoire des Probabilités et de la Statistique Electronic Journal for History of Probability and Statistics*. 2008. 4.

[72] Ester M, Kriegel HP, Sander J and Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, vol. 96. 1996 pp. 226–231.

[73] Day WH and Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*. 1984. 1(1) : 7–24.

[74] Sneath PH. The application of computers to taxonomy. *Microbiology*. 1957. 17(1) : 201–226.

[75] Sorenson T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*. 1948. 5(1-34) : 4–7.

[76] Sokal RR. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*. 1958. 38 : 1409–1438.

[77] Ward Jr JH. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. 1963. 58(301) : 236–244.

[78] McQuitty LL. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological measurement*. 1966. 26(4) : 825–831.

[79] Gower JC. A comparison of some methods of cluster analysis. *Biometrics*. 1967. pp. 623–637.

[80] Banfield JD and Raftery AE. Model-based gaussian and non-gaussian clustering. *Biometrics*. 1993. pp. 803–821.

[81] Fraley C, Raftery AE, *et al.* Model-based methods of classification : Using the mclust software in chemometrics. *Journal of Statistical Software*. 2007. 18(6) : 1–13.

[82] Celeux G and Govaert G. Gaussian parsimonious clustering models. *Pattern recognition*. 1995. 28(5) : 781–793.

[83] Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971. pp. 857–871.

[84] Genolini C, Pingault J, Driss T, Cote S, Tremblay RE, Vitaro F, *et al.* Kml3d : a non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine*. 2013. 109(1) : 104–111.

[85] Muthèn B and Shedden K. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*. 1999. 55(2) : 463–469.

[86] Proust C and Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer methods and programs in biomedicine*. 2005. 78(2) : 165–173.

[87] Caliński T and Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*. 1974. 3(1) : 1–27.

[88] Schwarz G, *et al.* Estimating the dimension of a model. *The annals of statistics*. 1978. 6(2) : 461–464.

[89] Fraley C and Raftery AE. How many clusters ? which clustering method ? answers via model-based cluster analysis. *The computer journal*. 1998. 41(8) : 578–588.

[90] Verbeke G and Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*. 1996. 91(433) : 217–221.

[91] Huang JZ and Shen H. Functional coefficient regression models for non-linear time series : A polynomial spline approach. *Scandinavian journal of statistics*. 2004. 31(4) : 515–534.

[92] McCulloch CE and Neuhaus JM. Generalized linear mixed models, Wiley Online Library2001.

[93] Genolini C, Alacoque X, Sentenac M and Arnaud C. kml and kml3d : R packages to cluster longitudinal data. *Journal of Statistical Software*. 2015. 65(1) : 1–34.

[94] Strenio JF, Weisberg HI and Bryk AS. Empirical bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*. 1983. pp. 71–86.

[95] Group BDW. Biomarkers and surrogate endpoints : Preferred definitions and conceptual framework. *Clinical Pharmacol & Therapeutics*. 2001. 69 : 89–95.

[96] Buyse M. Towards validation of statistically reliable biomarkers. *European Journal of Cancer Supplements*. 2007. 5(5) : 89–95.

[97] Roubertoux PL and de Vries PJ. From molecules to behavior : lessons from the study of rare genetic disorders. *Behavior genetics*. 2011. 41(3) : 341–348.

[98] Ferlini A, Scotton C and Novelli G. Biomarkers in rare diseases. *Public health genomics*. 2013. 16(6) : 313–321.

[99] Mahant N, McCusker E, Byth K, Graham S, *et al.* Huntington's disease clinical correlates of disability and progression. *Neurology*. 2003. 61(8) : 1085–1092.

[100] Rosenblatt A, Liang KY, Zhou H, Abbott M, Gourley L, Margolis R, *et al.* The association of CAG repeat length with clinical progression in huntington disease. *Neurology*. 2006. 66(7) : 1016–1020.

[101] Gusella JF, MacDonald ME and Lee JM. Genetic modifiers of huntington's disease. *Movement Disorders*. 2014. 29(11) : 1359–1365.

[102] Lotta T, Vidgren J, Tilgmann C, Ulmanen I, Melen K, Julkunen I, *et al.* Kinetics of human soluble and membrane-bound catechol o-methyltransferase : a revised mechanism and description of the thermolabile variant of the enzyme. *Biochemistry*. 1995. 34(13) : 4202–4210.

[103] Chen J, Lipska BK, Halim N, Ma QD, Matsumoto M, Melhem S, *et al.* Functional analysis of genetic variation in catechol-o-methyltransferase (comt) : effects on mrna, protein, and enzyme activity in postmortem human brain. *The American Journal of Human Genetics*. 2004. 75(5) : 807–821.

[104] Cools R and D'Esposito M. Inverted-u–shaped dopamine actions on human working memory and cognitive control. *Biological psychiatry*. 2011. 69(12) : e113–e125.

[105] Peinemann A, Schuller S, Pohl C, Jahn T, Weindl A and Kassubek J. Executive dysfunction in early stages of huntington's disease is associated with striatal and insular atrophy : a neuropsychological and voxel-based morphometric study. *Journal of the neurological sciences*. 2005. 239(1) : 11–19.

[106] Proust C, Jacqmin-Gadda H, Taylor JM, Ganiayre J and Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*. 2006. 62(4) : 1014–1024.

[107] Gennatas E, Cholfin J, Zhou J, Crawford R, Sasaki D, Karydas A, *et al.* Comt val158met genotype influences neurodegeneration within dopamine-innervated brain structures. *Neurology*. 2012. 78(21) : 1663–1669.

[108] Sargent DJ, Conley BA, Allegra C and Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005. 23(9) : 2020–2027.

[109] Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A and De Gramont A. Integrating biomarkers in clinical trials. 2011.

[110] Matsui S. Genomic biomarkers for personalized medicine : development and validation in clinical studies. *Computational and mathematical methods in medicine*. 2013. 2013.

[111] Eng KH. Randomized reverse marker strategy design for prospective biomarker validation. *Statistics in medicine*. 2014. 33(18) : 3089–3099.

[112] Simon R. Biomarker based clinical trial design. *Chin Clin Oncol*. 2014. 3 : 39.

[113] Matsui S, Choai Y and Nonaka T. Comparison of statistical analysis plans in randomize-all phase iii trials with a predictive biomarker. *Clinical Cancer Research*. 2014. 20(11) : 2820–2830.

[114] Freidlin B and Korn EL. Biomarker enrichment strategies : matching trial design to biomarker credentials. *Nature Reviews Clinical Oncology*. 2014. 11(2) : 81–90.

[115] Simon R and Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004. 10(20) : 6759–6763.

[116] Maitournam A and Simon R. On the efficiency of targeted clinical trials. *Statistics in medicine*. 2005. 24(3) : 329–339.

[117] Mandrekar SJ and Sargent DJ. Clinical trial designs for predictive biomarker validation : theoretical considerations and practical challenges. *Journal of Clinical Oncology*. 2009. 27(24) : 4027–4034.

[118] Freidlin B, McShane LM and Korn EL. Randomized clinical trials with biomarkers : design issues. *Journal of the National Cancer Institute*. 2010.

[119] Young K, Laird A and Zhou X. The efficiency of clinical trial designs for predictive biomarker validation. *Clinical Trials*. 2010.

[120] Landwehrmeyer GB, Dubois B, de Yébenes JG, Kremer B, Gaus W, Kraus PH, *et al.* Riluzole in huntington's disease : a 3-year, randomized controlled study. *Annals of neurology*. 2007. 62(3) : 262–272.

[121] Schramm C, Vial C, Bachoud-Lévi AC and Katsahian S. Clustering of longitudinal data by using an extended baseline : A new method for treatment efficacy clustering in longitudinal data. *Statistical methods in medical research*. 2015. p. 0962280215621591.

[122] Schramm C, Katsahian S, Youssov K, Démonet JF, Krystkowiak P, Supiot F, *et al.* How to capitalize on the retest effect in future trials on huntington's disease. *PloS one*. 2015. 10(12) : e0145842.

[123] James GM and Sugar CA. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*. 2003. 98(462) : 397–408.

[124] Ferreira L and Hitchcock DB. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*. 2009. 38(9) : 1925–1949.

[125] Lipkovich IA, Houston JP and Ahl J. Identifying patterns in treatment response profiles in acute bipolar mania : a cluster analysis approach. *BMC psychiatry*. 2008. 8(1) : 65.

[126] Tarpey T, Petkova E and Ogden RT. Profiling placebo responders by self-consistent partitioning of functional data. *Journal of the American Statistical Association*. 2003. 98(464) : 850–858.

[127] Baudelet C and Gallez B. Cluster analysis of bold fmri time series in tumors to study the heterogeneity of hemodynamic response to treatment. *Magnetic resonance in medicine*. 2003. 49(6) : 985–990.

[128] Yu L, Boyle PA, Segawa E, Leurgans S, Schneider JA, Wilson RS, *et al.* Residual decline in cognition after adjustment for common neuropathologic conditions. *Neuropsychology*. 2015. 29(3) : 335.

[129] Ortega H, Li H, Suruki R, Albers F, Gordon D and Yancey S. Cluster analysis and characterization of response to mepolizumab. a step closer to personalized medicine for patients with severe asthma. *Annals of the American Thoracic Society*. 2014. 11(7) : 1011–1017.

[130] Schreibmann E, Waller AF, Crocker I, Curran W and Fox T. Voxel clustering for quantifying pet-based treatment response assessment. *Medical physics*. 2013. 40(1) : 012401.

[131] McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, *et al.* Reporting recommendations for tumor marker prognostic studies (remark). *Journal of the National Cancer Institute*. 2005. 97(16) : 1180–1184.

[132] Cardoso F, Piccart-Gebhart M, Van't Veer L, Rutgers E, *et al.* The mindact trial : the first prospective clinical validation of a genomic tool. *Molecular oncology*. 2007. 1(3) : 246–251.

[133] Pepe MS, Janes H, Longton G, Leisenring W and Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*. 2004. 159(9) : 882–890.

[134] Dahlberg S and Liu PY. Prognostic factors in clinical trials. *Breast cancer research and treatment*. 1992. 22(3) : 193–196.

[135] Buyse M, Sargent DJ, Grothey A, Matheson A and De Gramont A. Biomarkers and surrogate end points—the challenge of statistical validation. *Nature reviews Clinical oncology*. 2010. 7(6) : 309–317.

[136] Schulz KF and Grimes DA. Generation of allocation sequences in randomised trials : chance, not choice. *The Lancet*. 2002. 359(9305) : 515–519.

[137] Kernan WN, Viscoli CM, Makuch RW, Brass LM and Horwitz RI. Stratified randomization for clinical trials. *Journal of clinical epidemiology*. 1999. 52(1) : 19–26.

[138] Beach ML and Meier P. Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials*. 1989. 10(4) : 161–175.

[139] Canner PL. Covariate adjustment of treatment effects in clinical trials. *Controlled clinical trials*. 1991. 12(3) : 359–366.

[140] Krijgsman O, Roepman P, Zwart W, Carroll JS, Tian S, de Snoo FA, *et al.* A diagnostic gene profile for molecular subtyping of breast cancer associated with treatment response. *Breast cancer research and treatment*. 2012. 133(1) : 37–47.

[141] Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*. 2008. 359(17) : 1757–1765.

[142] Holmes JA, Desmond PV and Thompson AJ. Redefining baseline demographics : the role of genetic testing in hepatitis c virus infection. *Clinics in liver disease*. 2011. 15(3) : 497–513.

[143] Jänne PA, Gurubhagavatula S, Yeap BY, Lucca J, Ostler P, Skarin AT, *et al.* Outcomes of patients with advanced non-small cell lung cancer treated with gefitinib (zd1839,'iressa') on an expanded access study. *Lung Cancer*. 2004. 44(2) : 221–230.

[144] Miller VA, Kris MG, Shah N, Patel J, Azzoli C, Gomez J, *et al.* Bronchioloalveolar pathologic subtype and smoking history predict sensitivity to gefitinib in advanced non–small-cell lung cancer. *Journal of Clinical Oncology*. 2004. 22(6) : 1103–1109.

[145] Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, *et al.* Egfr mutations in lung cancer : correlation with clinical response to gefitinib therapy. *Science*. 2004. 304(5676) : 1497–1500.

[146] Arns M, Drinkenburg WH, Fitzgerald PB and Kenemans JL. Neurophysiological predictors of non-response to rtms in depression. *Brain Stimulation*. 2012. 5(4) : 569–576.

[147] Mayberg HS, Brannan SK, Mahurin RK, Jerabek PA, Brickman JS, Tekell JL, *et al.* Cingulate function in depression : a potential predictor of treatment response. *Neuroreport*. 1997. 8(4) : 1057–1061.

[148] Noda A, Kraemer HC, Taylor JL, Schneider B, Ashford JW and Yesavage JA. Strategies to reduce site differences in multisite studies : a case study of alzheimer disease progression. *The American journal of geriatric psychiatry*. 2006. 14(11) : 931–938.

[149] Collie A, Maruff P, Darby DG and McSTEPHEN M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*. 2003. 9(03) : 419–428.

[150] Beglinger LJ, Gaydos B, Tangphao-Daniels O, Duff K, Kareken DA, Crawford J, *et al.* Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*. 2005. 20(4) : 517–529.

[151] Cooper D, Lacritz L, Weiner M, Rosenberg R and Cullum C. Category fluency in mild cognitive impairment : reduced effect of practice in test-retest conditions. *Alzheimer Disease & Associated Disorders*. 2004. 18(3) : 120–122.

[152] Cooper D, Epker M, Lacritz L, Weiner M, Rosenberg R, Honig L, *et al.* Effects of practice on category fluency in alzheimers disease*. *The Clinical Neuropsychologist*. 2001. 15(1) : 125–128.

[153] SNOWDEN J, Craufurd D, Griffiths H, Thompson J and Neary D. Longitudinal evaluation of cognitive disorder in huntington's disease. *Journal of the International Neuropsychological Society*. 2001. 7(01) : 33–44.

[154] Benedict RH and Zgaljardic DJ. Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*. 1998. 20(3) : 339–352.

[155] Katsnelson A. Momentum grows to make'personalized'medicine more'precise'. *Nature medicine*. 2013. 19(3) : 249–249.

[156] Trusheim MR, Berndt ER and Douglas FL. Stratified medicine : strategic and economic implications of combining drugs and clinical biomarkers. *Nature Reviews Drug Discovery*. 2007. 6(4) : 287–293.

[157] Hamburg MA and Collins FS. The path to personalized medicine. *New England Journal of Medicine*. 2010. 363(4) : 301–304.

[158] Golden C. Stroop colour and word test. *age*. 1978. 15 : 90.

[159] Mattis S. Mental status examination for organic mental syndrome in the elderly patient. *Geriatric psychiatry*. 1976. 11(77) : e121.

[160] Butters N, Wolfe J, Granholm E and Martone M. An assessment of verbal recall, recognition and fluency abilities in patients with huntington's disease. *Cortex*. 1986. 22(1) : 11–32.

[161] Cardebat D, Doyon B, Puel M, Goulet P and Joanette Y. [formal and semantic lexical evocation in normal subjects. performance and dynamics of production as a function of sex, age and educational level]. *Acta neurologica belgica*. 1989. 90(4) : 207–217.

[162] Zazzo R. Test des deux barrages, EAP1960.

[163] Brandt J. The hopkins verbal learning test : Development of a new memory test with six equivalent forms. *The Clinical Neuropsychologist*. 1991. 5(2) : 125–142.

[164] Rieu D, Bachoud-Lévi AC, Laurent A, Jurion E and Dalla Barba G. Adaptation française du «hopkins verbal learning test». *Revue neurologique*. 2006. 162(6) : 721–728.

[165] Reitan RM. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and motor skills*. 1958. 8(3) : 271–276.

[166] Kleinbaum DG and Klein M. Survival analysis, Springer1996.

[167] Cox DR. Partial likelihood. *Biometrika*. 1975. 62(2) : 269–276.

[168] Kelly PJ and Lim LLY. Survival analysis for recurrent event data : an application to childhood infectious diseases. *Statistics in medicine*. 2000. 19(1) : 13–33.

[169] Andersen PK and Gill RD. Cox's regression model for counting processes : a large sample study. *The annals of statistics*. 1982. pp. 1100–1120.

[170] Lee EW, Wei L, Amato DA and Leurgans S. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In Survival analysis : state of the art, pp. 237–247, Springer1992.

[171] Lin DY and Wei LJ. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*. 1989. 84(408) : 1074–1078.

[172] Box-Steffensmeier JM and De Boef S. Repeated events survival models : the conditional frailty model. *Statistics in medicine*. 2006. 25(20) : 3518–3533.

[173] Hougaard P. Frailty models for survival data. *Lifetime data analysis*. 1995. 1(3) : 255–273.

[174] Therneau TM, Grambsch PM and Pankratz VS. Penalized survival models and frailty. *Journal of computational and graphical statistics*. 2003. 12(1) : 156–175.

[175] Austin PC. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in medicine*. 2012. 31(29) : 3946–3958.

Résumé/Abstract

Résumé

La maladie de Huntington est neurodégénérative, génétique, rare, multifacette et de durée d'évolution longue, induisant une grande hétérogénéité sur la présentation et l'évolution de la maladie. Les biothérapies en cours d'essai sont réalisées sur des petits effectifs, avec un effet mesurable à long terme et hétérogène. Identifier des marqueurs d'évolution de la maladie et de réponse au traitement permettrait de mieux comprendre et d'améliorer les résultats des études de biothérapie dans la maladie de Huntington. Nous avons développé une méthode de clustering pour l'efficacité d'un traitement dans le cadre de données longitudinales afin de définir des répondeurs et non répondeurs au traitement. Notre méthode combine un modèle linéaire mixte à deux pentes et un algorithme de clustering classique (modèle le mélange, k-moyennes). Le modèle mixte génère des effets aléatoires, associés à la réponse au traitement, propres à chaque patient. L'algorithme de clustering permet de définir des sous-groupes selon la valeur des effets aléatoires. Notre méthode est robuste pour les petits effectifs. Trouver des sous-groupes de patients répondeurs permet de définir des marqueurs prédictifs de la réponse au traitement qui seront utilisés pour donner le traitement le mieux adapté à chaque patient. Nous avons discuté de l'intégration (i) des marqueurs prédictifs dans les plans expérimentaux des futurs essais cliniques, en évaluant leur impact sur la puissance de l'étude ; et (ii) des marqueurs pronostiques de l'évolution de la maladie, en étudiant le polymorphisme COMT comme marqueur pronostique du déclin cognitif des patients atteints de la maladie de Huntington. Enfin, nous avons évalué l'effet d'apprentissage des tests neuropsychologiques mesurant les capacités cognitives, et montré comment une double évaluation à l'inclusion dans un essai clinique permettait de s'en affranchir quand le critère de jugement principal est le déclin cognitif.

Mots clefs : Clustering ; données longitudinales ; plans expérimentaux ; médecine stratifiée ; retest ; maladie de Huntington

Abstract

Integration of predictive factors of treatment effect in design and analyse of clinical trials with small sample size : application on Huntington's disease.

Huntington's disease is neurodegenerative, genetic, rare, multifaceted and has a long evolution, inducing heterogeneity of conditions and progression of the disease. Current biotherapy trials are performed on small samples of patients, with a treatment effect measurable in the long-term that is heterogeneous. Identifying markers of the disease progression and of the treatment response may help to better understand and improve results of biotherapy studies in Huntington's disease. We have developed a clustering method for the treatment efficacy in the case of longitudinal data in order to identify treatment responders and nonresponders. Our method combines a linear mixed model with two slopes and a classical clustering algorithm (model-based, k-means). The mixed model generates random effects associated with treatment response, specific to each patient. The clustering algorithm is used to define subgroups according to the value of the random effects. Our method is robust in case of small samples. Finding subgroups of responders may help to define predictive markers of treatment response which will be used to give the most appropriate treatment for each patient. We discussed integration of (i) the predictive markers in study design of future clinical trials, assessing their impact on the power of the study ; and (ii) the prognostic markers of disease progression by studying the COMT polymorphism as a prognostic marker of cognitive decline in Huntington's disease. Finally, we evaluated the learning effect of neuropsychological tasks measuring cognitive abilities, and showed how a double baseline in a clinical trial could take it into account when the primary outcome is the cognitive decline.

Keywords : Clustering ; longitudinal data ; designs ; stratified medicine ; retest ; Huntington's disease