Item Response Theory in the Neurodegenerative Disease Data Analysis

Wenjia Wang

DOCTEURE DE

L'UNIVERSITÉ DE BORDEAUX

École Doctorale Sociétés, Politique, Santé Publique

Spécialité Santé Publique, option Biostatistique

Item Response Theory in the Neurodegenerative Disease Data Analysis

---

Par Wenjia WANG

Sous la direction de Daniel COMMENGES

Co-directeur : Mickaël GUEDJ

Soutenu le 21/06/2017

Membres du jury

Hélène JACQUE-GADDA	Directrice de Recherche, INSERM, Bordeaux	Présidente
Grégory NUEL	Professeur, UPMC, Paris	Rapporteur
Véronique SEBILLE-RIVAIN	Directrice de Recherche, INSERM, Nante	Rapporteur
Daniel COMMENGES	Directeur de Recherche, INSERM, Bordeaux	Directeur de thèse
Mickaël GUEDJ	Ph.D, Pharnext, Paris	Co-directeur de thèse
Sandrine ANDRIEU	Directrice de Recherche, INSERM, Toulouse	Examinateur

Théorie de la réponse d'item dans l'analyse des données sur les maladies neurodégénératives

Les maladies neurodégénératives, telles que la maladie d'Alzheimer (AD) et Charcot Marie Tooth (CMT), sont des maladies complexes. Leurs mécanismes pathologiques ne sont toujours pas bien compris et les progrès dans la recherche et le développement de nouvelles thérapies potentielles modifiant la maladie sont lents. Les données catégorielles, comme les échelles de notation et les données sur les études d'association génomique (GWAS), sont largement utilisées dans les maladies neurodégénératives dans le diagnostic, la prédiction et le suivi de la progression. Il est important de comprendre et d'interpréter ces données correctement si nous voulons améliorer la recherche sur les maladies neurodégénératives. Le but de cette thèse est d'utiliser la théorie psychométrique moderne: théorie de la réponse d'item pour analyser ces données catégoriques afin de mieux comprendre les maladies neurodégénératives et de faciliter la recherche de médicaments correspondante. Tout d'abord, nous avons appliqué l'analyse de Rasch afin d'évaluer la validité du score de neuropathie Charcot-Marie-Tooth (CMTNS), un critère important d'évaluation principal pour les essais cliniques de la maladie de CMT. Nous avons ensuite adapté le modèle Rasch à l'analyse des associations génétiques pour identifier les gènes associés à la maladie d'Alzheimer. Cette méthode résume les génotypes catégoriques de plusieurs marqueurs génétiques tels que les polymorphisme nucléotidique (SNPs) en un seul score génétique. Enfin, nous avons calculé l'information mutuelle basée sur la théorie de réponse d'item pour sélectionner les items sensibles dans ADAS-cog, une mesure de fonctionnement cognitif la plus utilisées dans les études de la maladie d'Alzheimer, afin de mieux évaluer le progrès de la maladie.

Mots clés : Maladie neurodegenerative; echelle de notation; données categoriques; theorie de la réponse d'item; Modèle Rasch; Analyse Rasch; GWAS; test d'association génétiques; Maladie d'Alzheimers; Maladie Charcot-Marie-Tooth; CMTNS; information mutuelle; ; ADAS-cog

Item Response Theory in the Neurodegenerative Disease Data Analysis

Neurodegenerative diseases, such as Alzheimer's disease (AD) and Charcot Marie Tooth (CMT), are complex diseases. Their pathological mechanisms are still not well understood, and the progress in the research and development of new potential disease-modifying therapies is slow. Categorical data like rating scales and Genome-Wide Association Studies (GWAS) data are widely utilized in the neurodegenerative diseases in the diagnosis, prediction and progression monitor. It is important to understand and interpret these data correctly if we want to improve the disease research. The purpose of this thesis is to use the modern psychometric Item Response Theory to analyze these categorical data for better understanding the neurodegenerative diseases and facilitating the corresponding drug research. First, we applied the Rasch analysis in order to assess the validity of the Charcot-Marie-Tooth Neuropathy Score (CMTNS), a main endpoint for the CMT disease clinical trials. We then adapted the Rasch model to the analysis of genetic associations and used to identify genes associated with Alzheimer's disease by summarizing the categorical genotypes of several genetic markers such as Single Nucleotide Polymorphisms (SNPs) into one genetic score. Finally, to select sensitive items in the most used psychometrical tests for Alzheimer's disease, we calculated the mutual information based on the item response model to evaluate the sensitivity of each item on the ADAS-cog scale.

Keywords : Neurodegenerative disease; Rating scale; categorical data; Item response theory; Rasch Model; Rasch analysis; GWAS; Gene-based association test; Alzheimer's disease; Charcot-Marie-Tooth disease; CMTNS; mutual information; ADAS-cog

Discipline : Santé publique - option : Biostatistiques

Laboratoire : Unité INSERM U1219 - Université de Bordeaux - ISPED 146 rue Léo Saignat 33076 Bordeaux, FRANCE

To my dear parents

Contents

Acknowledgement

During the years of my thesis, I always felt grateful to have the chance to have these people around me. I would like to express my most sincere gratitude to them here.

First, I would like to thank both of my directors. Mickaël Guedj, since the first day I worked with you, you have taught me countless things. From writing R program to scientific spirit, from paying attention to details to paying attention to other's need. You are always available for discussion and have given my good advice. For a foreign student, there are plentiful difficulties in study and in culture to overcome. All the time, you try to think into my place and kindly help me with your warm smile. I will not forget your guidance, encouragement and patience. Daniel Commenges, although I haven't stay in Bordeaux often enough, you still try your best to advise me. In those days when I had the chance to share your office, your in-depth knowledge of statistics, your rigorous attitude of scholarship and also your humour sense set me a great example. I enjoy the talks about the science and music with you. You have provided me inspirations for my research that I would like to continue working on.

I would also like to thank the president of jury Hélène Jacque-Gadda. The discussions for hours with you and Cécile Proust-Lima about the research were so inspiring and helpful. I feel so grateful for the time and concentration you render, even though I am not your own student. You have also reviewed my articles so carefully and your questions made me try to think in deep. I really appreciate your warm heart and enthusiasm for research.

Thanks to the examiners and reporters of jury. Grégory Nuel, as the director of Mickaël in his thesis, that you can accept review my thesis makes me feel happy and honored. Véronique Sébille, I'm so grateful that you can accept to take the time to review my thesis. Last but not least, thank you Sandrine Andrieu for acceptance to be a member of jury!

Thanks, Daniel Cohen, without you, there would have been no Pharnext and consequently no this thesis CIFRE. It's you who makes me see potentials on myself and have the chance to apply them. How your encouragements inspired me! You are never tired of the endless work and endless jokes. Your great optimism and empathy will always be my model.

The Biostatistics and Bioinformatics team of Pharnext, it is such a pleasure to work with you together, no matter you have already left (Jonas, thanks for your answering all my questions so patiently, it was so helpful! Sory Traoré, you have brought so much fun into the office and it was always nice to talk to you. Claudia Sabizabal, your works provided a solid base to my researches. Wish you have a brilliant career in the US!), or you are still here but in another office now (Jan Bouaziz, your jokes are cold but your heart is warm, thanks for the professional and personal help. Julie, you are always cheerful and caring, even when you are down. Hope I can learn from your courage and optimism. Peter Schmitt, you are calm but you show your enthusiasm and perfectionism on your work, thanks a lot for your help! Julien Laffaire, although I always fail to get your jokes, still I feel grateful for your attentive and kindness behind. Fabrice Gilbert, it's always so relaxing to talk with you, thank you all the great advice from the internet to music! Philippe Rinaudo, you have impressed me with your patience and passion when you were our TA. It's an honor to work with you).

I am also grateful to the other colleagues in Pharnext, Marina Ossokina and Esther Graudens who share the office with me and three years of accompaniment, the secretaries Linda Pondicq-Cassou and Catherine Barret who help me out from all the administrative complications, Niall Murphy and Viviane Bertrand who has carefully reviewed my article, Serguei Nabirotchkin for his scientific advice, Magali Marcadé, Natalie Cholet, Chantal Covos, Caroline Mertens, Agnes Daoust…every one in Pharnext. It's lucky to work in such a friendly ambiance.

Personally, I would like to thank my friends, Nan Nan and Xiaoshan Liu. You have accompanied me during all these years of my thesis. You cheered me up when I was in confusion and share the happiness of the life with me. Stephane and Chloe, I talk with you in a language which is not my mother-tongue, but it doesn't stop that we understand and support each other! For those who are far from me now, Yuting, Qian, Lizhen, Baobao… We never stop our contact, even though we haven't met for years. Thanks for all your encouragement!

In the end, I would like to thank my dearest parents. I know it was really not easy for you to let your only daughter go aboard for years. Still, you always try to understand what I am doing and encourage me to be what I want to be. It's always full of pleasure to talk with you. You are not only my parents, but also my best friends. So lucky and proud to be your daughter!

Scientific Production

Articles

Thesis publications

♦ Wenjia Wang, Jonas Mandel, Jan Bouaziz, Daniel Commenges, Serguei Nabirotchkine, Ilya Chumakov, Daniel Cohen, Mickaël Guedj, A Multi-Marker Genetic Association Test Based on the Rasch Model Applied to Alzheimer's Disease, *Plos One*, 2015

♦ Wenjia Wang, Mickaël Guedj, Viviane Bertrand, Julie Foucquier, Elisabeth Jouve, Daniel Commenges, Cécile Proust-Lima, Niall P. Murphy, Olivier Blin, Laurent Magy, Daniel Cohen, Shahram Attarian, A Rasch Analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a Cohort of Charcot-Marie-Tooth Type 1A Patients, *Plos One*, 2017

♦ Wenjia Wang, Mickaël Guedj, Helene Jacqmin-Gadda, Cecile Proust-Lima, Daniel Commenges, Selection of items as sensitive clinical markers for MCI population from the ADAS–Cog with the IRT-based mutual information. In submission.

Collaborations

♦ Qiang Luo, Qiang Chen, Tomáš Paus, Wenjia Wang, Sylvane Desrivières, Christine Macare, Tianye Jia, Gabriel H. Robert, Mickaël Guedj, Jing Cui, Joseph H. Callicott, Venkata S. Mattay, Zdenka Pausova, Karen F. Berman, Daniel R. Weinberger, Gunter Schuman, Jianfeng Feng, A risk marker for schizophrenia mediates a gene-by-environment interaction on psychotic symptom by influencing putamen volume across the life span. In submission

Communications

Oral communications at conferences

♦ Wenjia Wang, Mickaël Guedj, A New Gene Based test of Association Using Extended Rasch Models. Statistical Methods for Post-Genomic Data, Paris, France, 2014

♦ Wenjia Wang, Multi-Marker Genetic Association Test Based on the Rasch Model Provides New Insights into Genetic of Alzheimer's Disease Symposium Advances in Systems Biology in Neurosciences, Geneva, Swiss, 2015

♦ Wenjia Wang, A multi-marker genetic association test based on the Rasch model provides new insights into genetics of Alzheimer's disease, 43nd European Mathematical Genetics Meeting, Brest, France, 2015

♦ Wenjia Wang, Mutual Information in Item-Response-Theory (IRT) Models and its Application to Alzheimer Disease, 7nd Rencontres des Jeunes Statisticiens, Porquerolles, France, 2017

Written communications (posters) at conferences

♦ Wenjia Wang, A New Gene-Based Test of Association Using Rasch Models, 42nd European Mathematical Genetics Meeting, Cologne, Germany, 2014

Notations and abbreviations

Notations

θ: Person ability parameter

β: Item difficulty parameter

D: Scaling factor

α: Discrimination parameter

c: Pseudo-guessing parameter

lnL: Log likelihood function

H: Entropy

I: Mutual information

Abbreviations

1-PL: One-Parameter Model

2-PL: Two-Parameter Model

3-PL: Three-Parameter Model

AD: Alzheimer's Disease

ADNI: Alzheimer's Disease Neuroimaging Initiative

ADAS-cog: Alzheimer's Disease Assessment Scale-cognitive subscale

CCC: Category Characteristic Curve

CMT: Charcot Marie Tooth Disease

CMT1A: Charcot Marie Tooth Type 1A Disease

CMTNS: Charcot-Marie-Tooth Neuropathy Score

CTT: Classical Test Theory

EOAD: Early-Onset Alzheimer's Disease

GRM: Graded Response Theory

GWAS: Genome Wide Association Study

ICC: Item Characteristic Curve

IRT: Item Response Function

IRT: Item Response Theory

LOAD: Late-Onset Alzheimer's Disease

MCI: Mild Cognitive Impairment

MI: Mutual information

PCM: Partial Credit Model

RR: Relative risk

RSM: Rating Scale Model

SNP: Single Nucleotide Polymorphism

Table of Figures

摘要（Abstract in Chinese)

神经退行性疾病包括阿兹海默症和腓骨肌萎缩症是一类复杂疾病。它们的病理机制仍然未被很好地理解，并且开发新疗法的研究和开发进展缓慢。分类数据如评级量表和全基因组关联研究数据被广泛应用于神经退行性疾病的诊断，预测和进展监测。如果我们想改善疾病研究，正确理解和解释这些数据是很重要的。本论文的目的是使用现代心理测量理论：项目反应理论来分析这些分类数据，以更好地了解神经退行性疾病和促进相应的药物研究。首先，我们应用 Rasch 分析，以评估腓骨肌萎缩症严重程度的主流评分量表和临床试验的主要终点腓骨肌萎缩症神经病变量表的有效性。然后，我们将 Rasch 模型用于遗传关联分析。这种方法中通过将一个基因中多个遗传标记如单核苷酸多态性的分类数据归纳为一个遗传分数来鉴定与阿兹海默症相关的基因。最后，为了在阿兹海默症痴呆评定量表中选择其中对疾病严重程度变化敏感的项目，我们基于项目响应模型计算互信息熵，以评估每个项目的敏感性。

关键词：神经退行性疾病;评分表; 全基因组关联分析; 分类数据; 项目反应理论; Rasch 模型; Rasch 分析; 基因关联测试; 阿兹海默症; 腓骨肌萎缩; 腓骨肌萎缩症神经病变量表; 互信息熵; 阿兹海默症痴呆评定量表

Résumé substantiel

Les maladies neurodégénératives telles que la maladie de Charcot-Marie-Tooth (CMT) et la maladie d'Alzheimer (AD) sont un groupe hétérogène de troubles qui se caractérisent par une dégénérescence progressive de la structure et de la fonction du système nerveux central ou du système nerveux périphérique. Ces maladies causent des problèmes de mouvement ou de fonctionnement mental chez les patients et peuvent être graves ou mettant la vie en danger. La plupart d'entre eux n'ont pas de remède efficace à jour.

L'analyse de ces maladies complexes a conduit au développement de nombreuses nouvelles méthodes technologiques, informatiques et analytiques qui visent à comprendre les mécanismes pathologiques de ces maladies. Différents types de données, tels que les SNPs et les scores des échelles de notation, sont utilisés dans le diagnostic, la prédiction et le suivi de la progression des maladies et des essais cliniques. Pour identifier les gènes qui ont un rôle dans un réseau de la maladie, les études d'association pangénomique (GWAS) qui scannent de grandes portions du génome afin de détecter les marqueurs génétiques sont nécessaires. GWAS donnent généralement des résultats au niveau de polymorphisme nucléotidique (SNPs). Cependant, la majorité des SNPs présentent des effets modestes et n'expliquent souvent qu'une petite partie de la variance ou de l'héritabilité du phénotype observé. Par conséquent, des modèles qui mesurent l'association combinée de multiples SNPs sont nécessaires. Pour évaluer la progression ou la gravité de la maladie, de nombreuses échelles de notation sont utilisées dans les essais cliniques des maladies neurodégénératives. Plusieurs questions concernant ces échelles émergent dans ce processus, y compris la validation et la sensibilité.

Les données SNP et les données des échelles de notation pourraient être considérées comme des données catégoriques. En général, la théorie des tests classiques est appliquée pour analyser les échelles avec des items catégorisés par lesquels les scores des items sont additionnés pour donner un score total. Comme alternative à la théorie de test classique, la théorie de réponse d'item (IRT) a été largement appliquée à l'analyse catégorielle des données depuis des années. Au lieu d'utiliser les scores directement additionnés d'un test, la théorie de la réponse de l'élément évalue la capacité de la personne par les difficultés de l'élément et ses réponses à ces items.

Cette thèse a été conçue sur la base d'une CIFRE (Conventions Industrielles de Formation par le REcherche) en collaboration entre la société Pharnext et le Centre de Santé Publique de l'Université de Bordeaux. L'objectif est de fournir des solutions pratiques et de nouvelles méthodes utilisant l'IRT pour répondre aux questions soulevées dans les études sur les maladies neurodégénératives.

L'IRT fournissent un cadre statistique d'analyse de mesure qui peut être utilisé pour approximer les fonctions de densité de probabilité dans la mesure. Il pourrait être appliqué à une échelle de notation lorsque les éléments satisfont à l'hypothèse d'unidimensionalité et d'indépendance locale. La probabilité que la réponse d'une personne à un item dans une échelle de notation peut donc être modélisée en utilisant une fonction logistique avec les paramètres de la capacité de la personne et des caractères des items.

Pour les items ayant seulement deux catégories, il existe trois principaux types de modèles IRT. Le modèle à un paramètre (1-PL) suppose que tous les éléments rapportent le trait latent de manière égale et les éléments ne varient que dans la difficulté. Le modèle à deux paramètres (2-PL) étend le 1-PL en estimant un paramètre de discrimination d'élément illustrant la capacité d'un élément à discriminer entre des traits contigus proches du point d'inflexion. Le modèle à trois paramètres (3-PL) étend le 2-PL en incluant un paramètre de deviner, qui ajuste pour l'impact du hasard sur les scores observés. Pour les articles comportant plus de deux catégories, plusieurs extensions IRT pourraient être adaptées, telles que le modèle d'échelle d'évaluation, le modèle de crédit partiel et le modèle de réponse graduée. Le modèle Rasch est un cas particulier de l'IRT. Bien que similaire à la 1-PL qui suppose l'égalité de la discrimination article, le modèle Rasch est plus une analyse confirmatoire qui met l'accent sur la primauté que le modèle corresponde aux données observées. Une fois que les données satisfont aux hypothèses du modèle de Rasch, une série de tests tels que l'ajustement et la cohérence de l'élément pourraient être appliqués pour évaluer l'échelle.

Cette thèse repose sur trois grands projets utilisant l'IRT et le modèle de Rasch dans différents aspects de l'analyse des données sur les maladies neurodégénératives et sont présentés ci-dessous.

Comme critère d'efficacité principal pour les essais cliniques de CMT, le score de neuropathie de Charcot-Marie-Tooth (CMTNS) est interrogé pour sa sensibilité au changement et ses propriétés

psychométriques sont encore discutées. Une méthode bien acceptée pour fournir la preuve de la validation de l'échelle sur la maladie est d'effectuer une analyse de modèle Rasch. Dans une première partie de cette thèse, nous avons utilisé l'analyse de Rasch pour évaluer les propriétés psychométriques du CMTNS avec une cohorte française de patient CMT1A. Nous avons d'abord testé les trois hypothèses de base: l'unidimensionalité, l'invariance et l'indépendance locale des scores de CMTNS. Une fois que ces hypothèses sont remplies, nous pourrions alors examiner la qualité générale de l'ajustement, la fiabilité, et la cohérence des items dans cette échelle. Les résultats de l'analyse nous ont permis de constater que le CMTNS est un mesure valide pour CMT1A. La plupart des items dans le CMTNS adaptent bien le modèle, sauf que deux items ont montré un overfit et 3 avaient des catégories désordonnées. Les résultats ont également souligné une limitation de la CMTNS est que les items sont plus adaptés pour évaluer les formes modérées à sévères de la maladie. Un perfectionnement plus poussé du CMTNS, comme l'ajout d'articles et/ou de catégories pour des évaluations de gravité modérée à modérée, est certainement à prendre en considération.

La détermination des gènes associés à la maladie pourrait faciliter la compréhension du mécanisme pathologique et le développement du traitement. Les génotypes des SNPs peuvent être codés en 0, 1 et 2 et donc ils pourraient être considérés comme une échelle de notation avec des items polytomes. Par conséquent, nous avons utilisé le modèle Rasch dans la deuxième partie du travail comme un test d'association génétique multi-marqueurs pour identifier les gènes associés à la maladie d'Alzheimer. Pour chaque gène, le modèle Rasch fournit une estimation de l'emplacement des individus sur le continuum de traits latent (dégrée d'association avec la maladie). En comparant ces emplacements du groupe de cas et du groupe control, nous avons pu évaluer l'association entre ce gène et la maladie. Nous avons conçu une série de simulations pour comparer cette méthode avec les quatre autres tests d'association existants. En comparant le taux de faux positifs et la puissance, nous avons trouvé que l'approche proposée a montré de bonnes performances. Ensuite, ce test d'association fondé sur le modèle Rasch a été appliqué aux données GWAS dans l'étude ADNI pour trouver les gènes associés à la maladie d'Alzheimer. Dans les gènes sélectionnés, plusieurs peuvent être fonctionnellement liés à la maladie. Une analyse de la voie de ces gènes met également en évidence le métabolisme du cholestérol qui joue un rôle clé dans la pathogenèse AD. De plus, ces éléments peuvent être intégrés dans un réseau de

signalisation hypothétique potentiellement ciblé par un médicament qui a montré une efficacité sur les modèles de maladie.

Dans la troisième partie de la thèse, nous avons exploré la combinaison de l'IRT et la théorie de l'information. Un item qui est plus dépendant du trait latent estimé par l'échelle a une plus grande possibilité de correspondre au changement de gravité de la maladie et donc pourrais être un marqueur sensible pour certaine population. L'information mutuelle fournit une estimation générale des dépendances en quantifiant la dépendance entre la distribution conjointe de deux variables. La distribution de la densité de probabilité des items pourrait être estimée par l'IRT. Le but de cette étude est de sélectionner les items sensibles dans ADAS-cog, une mesure de fonctionnement cognitif la plus utilisées dans les études de la maladie d'Alzheimer, afin de mieux évaluer le progrès de la maladie. Dans cette étude, nous avons calculé l'information mutuelle basée sur l'IRT pour chaque item dans l'ADAS-cog en utilisant les données des patients en MCI dans l'étude ADNI et l'a comparée avec d'autres statistiques fondées sur l'IRT. L'information mutuelle des items est mieux corrélée à l'évolution de sévérité des patients sur les données de suivi de deux ans. Dans l'ADAS-cog, les items Word Recall, Word Recognition et Delayed Word Recall ont des informations mutuelles plus élevées. Leur score composite a montré un taux de changement plus élevé par rapport à des scores composites des autres sous-échelles ADAS-cog. Cette étude indique que l'information mutuelle basée sur l'IRT pourrait être un critère de la sensibilité des items.

Preface

Neurodegenerative diseases, such as Charcot-Marie-Tooth disease and Alzheimer's disease, are a heterogeneous group of disorders that are characterized by the progressive degeneration of the structure and function of the central nervous system or peripheral nervous system. These diseases cause problems with movement or mental functioning in patients and can be serious and even life-threatening. Most of them have no effective cure up to date.

Neurodegenerative diseases are the major focus of Pharnext, a biopharmaceutical company founded in April 2007 by Professor Daniel Cohen and collaborators. Pharnext's mission is to discover and develop new therapeutic solutions for the severe orphan (relative rare) and common neurodegenerative diseases and companion tests for unmet medical needs.

The R&D approach of Pharnext consists in combining mini-doses of several drugs already approved by healthcare authorities for other diseases that are unrelated from a clinical viewpoint but linked in regarding the underlying biological networks. Pharnext's core expertise is based on reconstructing extensive disease networks using complex and extensive genomic data to identify the thousands of molecules possibly involved in a disease. This type of biological information is considered to be the "missing link" in pharmaceutical research. From disease molecular networks, Pharnext deduces synergistic combinations of drugs already approved but for unrelated indications. These novel combinations of drugs are called Pleodrugs. The classical R&D approaches to find therapeutic molecules are usually based on the "one drug, one disease" paradigm under which a single drug is used to treat a single yet often multifactorial diseases which are the result of the combined effect of several genes and the environment. On the other hand, Pleodrugs are capable of restoring the molecular pathways perturbed in diseases and addressing the shortcomings of the standard R&D approach that has shown its limits in terms of efficacy and safety. The novel strategy of Pharnext allows targeting several molecular 'nodes' in a disease-perturbed pathway and thus helps to increase the treatment efficacy and safety.

Pharnext has two lead products in clinical development: PXT3003 is currently in an international Phase 3 trial for the treatment of Charcot-Marie-Tooth disease type 1A in Europe and the United States. PXT864 has generated positive Phase 2 results in Alzheimer's disease.

Neurodegenerative diseases are often multifactorial diseases. The analysis of such complex diseases has led to the development of many new technological, computational and analytical methods that aim to understand the underlying complex mechanisms of these diseases. Various types of data, such as SNPs and scores of rating scales, are employed by Pharnext in the neurodegenerative diseases research and the clinical trials. To ensure that the neurodegenerative disease studies in Pharnext could provide a proper interpretation of their findings, the development of statistical methods is critical.

To identify genes which have a role in a disease network, Genome Wide Association Studies that screen large portions of the genome in order to detect genetic markers are necessary. GWAS generally yield results at the SNP-level. However, the majority of SNPs show modest effects and often explain only a small part of the variance or heritability of observed phenotypes. Therefore, models that measure the combined association of multiple SNPs are needed. To evaluate the progression of patients taking the medications, numerous rating scales are utilized in the neurodegenerative disease clinical trials. Several issues concerning these scales emerge in this process. First, some scales, such as the Charcot-Marie-Tooth Neuropathy Score developed as a man efficacy endpoint for clinical trials of Charcot-Marie-Tooth disease type 1A disease are suggested to be insensitive to disease progression. Second, some neurodegenerative studies have included multiple rating scales which measure different facets of the disease progression. It is of importance to examine the validation of the scale for the disease measurement and also to select sensitive items in existed tests to measure the disease progression in a certain group of patients.

SNP data and rating scales data could be considered as categorical data. Typically, the Classical Test Theory is applied to analyze the scales with categorized items whereby the item scores are summed to give a total score. As an alternative to the classical test theory, Item Response Theory has been widely applied to categorical data analysis. Instead of using the directly summed scores of a test, item response theory assesses the person ability by the item difficulties and their answers to items. A strict model of item response theory, Rasch model, provides a framework to evaluate a

rating scale. The item response theory may provide valuable indications on the data analysis in neurodegenerative disease research and helps interpret the finding in these studies.

To fully figure out the targeted problems and appropriately apply the item response theory, it is also important to collaborate with research groups and laboratories which are experienced in statistics and disease research.

The Bordeaux Population Health is a research centre belonging to Bordeaux University. It is dedicated to developing statistical methods for analysis of cohort data. The Biostatistics team develops new models and methods for epidemiology. Their emphasis is placed on dynamic models. The spectrum goes from theoretical to applied research. The main applications are in the epidemiology of Alzheimer's disease, AIDS (acquired immune deficiency syndrome), and cancers. The team will also develop an activity in statistical genetics.

My PhD thesis was designed on the basis of a CIFRE (Conventions Industrielles de Formation par la REcherche) in collaboration between the company Pharnext and the Bordeaux Population Health center. This collaboration focuses on the practical research needs of Pharnext and methodological developments. The objective is to provide practical solutions and new methods that answer the questions raised in the neurodegenerative disease studies. In this thesis, Item Response Theory is applied to the data of Neurodegenerative diseases in different ways. After a comparison with classical test theory and detailed description of Item response theory in the introduction, the Rasch model is served as a gene-based association test to identify genes that are associated with Alzheimer's disease. Then we use the Rasch analysis as a statistical framework of scale validation to examine a rating scale for the Charcot-Marie-Tooth Disease evaluation which is called CMTNS. Finally, the Item Response Theory based mutual information is applied to select sensitive items in the scale for the evaluation of cognitive functions in the Alzheimer's disease.

Chapter 1

Introduction

1.1. Neurodegenerative Disease

Neurodegenerative diseases are complex and often multifactorial. Main risk factors include certain genetic polymorphisms and aging. Other possible causes may include gender, poor education, endocrine conditions, oxidative stress, inflammation, stroke, hypertension, diabetes, smoking, head trauma, depression, infection, tumors, vitamin deficiencies, immune and metabolic conditions and chemical exposure (Brown et al., 2005).

The pathological mechanisms of neurodegenerative diseases remain not well understood; their phenotypes are largely symptom-based and not well defined. There is considerable overlap between the various neurodegenerative diseases (Brown et al., 2005) which increases the difficulties of identifying and evaluating diseases. However, many similarities are found in neurodegenerative diseases to relate them on a sub-cellular level. First, many neurodegenerative diseases are caused by genetic mutations. Although most of these mutations locate in completely unrelated genes, they still share some common features, such as a repeat of the CAG nucleotide triplet (Thompson, 2008). Second, aggregation of misfolded proteins happens in several neurodegenerative diseases, such as the aggregation of hyperphosphorylated tau protein that are known as a primary marker of Alzheimer's disease (Soto, 2003). These diseases also have some common intracellular mechanisms including the protein degradation pathways and mitochondrial dysfunction (Rubinsztein, 2006) as well as induced cell death (Bredesen et al., 2006). These relations between neurodegenerative diseases offer the hope of finding an effective treatment that could ameliorate many diseases simultaneously.

Up to date, there are no therapies available to cure neurodegeneration. The existing medications can only alleviate symptoms and help to improve patients' quality of life. For example,

memantine (Tariot et al., 2004) and donepezil (Birks and Harvey, 2006) can slow the progression of dementia symptoms in some people with Alzheimer's disease.

1.1.1. Alzheimer's disease

Epidemiology

Alzheimer's disease (AD) is the most common neurodegenerative disorder. According to the World Alzheimer report, there are 47.5 million people having dementia and 60% to 70% of dementia was caused by AD, and the number is expected to reach 65.7 million in 2030 and 115.4 million in 2050 (Weiner et al., 2012). The most common early symptom is the difficulty in remembering recent events (Burns and Iliffe, 2009). As the disease advances, progressive deterioration of cognitive functions appears, involving memory, reason, judgment and orientation. This disease can be characterized by brain atrophy reflecting neuronal and synaptic loss and the presence of amyloid plaques and neurofibrillary tangles. According to the age at onset, two main types of AD are differentiated: Early-Onset AD (EOAD) which generally appears before the age of 65 and Late-Onset AD (LOAD) appears after the age of 65 (Rogaeva, 2002). EOAD accounts for less than 10% of the AD population whereas LOAD accounts fo more than 90% of the AD population and has a complex etiology based on genetic and environmental factors.

The progression of Alzheimer's disease can be divided into four stages. The stages are sometimes overlapped. However, the Mini–Mental State Examination (MMSE) which is a 30-point questionnaire is often used in clinical to measure cognitive impairment and estimate the stage of dementia (Detecting Dementia with the Mini-Mental State Examination (MMSE) in Highly Educated Individuals).

Mild Cognitive Impairment (MCI). It is a condition in which someone has minor problems with their mental abilities such as short-term memory loss. MCI is a transitional stage between normal aging and dementia which frequently seen as a prodromal stage of Alzheimer's disease. People with MCI are more likely to go on to develop dementia (Grundman et al., 2004).

Mild stage (early stage). People who are diagnosed as AD show increasingly cognitive impairment. Difficulties with language, executive functions or execution of movements are more

prominent than memory problems. In this stage, a person may still be able to act independently, but their difficulties could be noticed such as Having greater difficulty performing tasks in social or work settings (Förstl and Kurz, 1999).

Moderate stage (middle stage). This stage can last very long for many years. People in this stage lose their independence gradually. Their symptoms such as forgetfulness of events or of one's personal history may be noticed by others, but patients may still remember significant details about their life.

Severe stage (Late stage). In this stage, individuals lose the ability to respond to their environment, to carry on a conversation and, eventually, to control movement.

There is an urgent need of disease-modifying treatments to slow or halt AD pathology progression on the population at risk for development of cognitive decline and dementia. Studies show that cognitive reserve, physical activity and exercise, midlife obesity, alcohol intake, and smoking are the most important modifiable risk factors for AD (Ballard et al., 2011). Nevertheless, AD pathogenic mechanisms are still unclear, and the disease remains a condition without cure. Several competing hypotheses try to explain the cause of the disease: the amyloid hypothesis supposes that extracellular amyloid beta (A_β) deposits are the fundamental cause of the disease (Hardy and Allsop, 1991); the tau hypothesis proposes that tau protein abnormalities initiate the disease cascade (Mudher and Lovestone, 2002); the genetic heritability of AD (and memory components thereof), based on reviews of twin and family studies, range from 49% to 79% (Gatz et al., 2006; Wilson et al., 2011), the *APOE* ε4 allele is the strongest known genetic risk factor for AD. The risk increased a two- to three- fold in people with one *APOE* ε4 allele and about 12-fold in those with two alleles (Trzepacz et al., 2014). However, it is neither sufficient nor necessary to explain all occurrences of disease. The dominant mutations in the genes encoding amyloid precursor protein (APP) and presenilin 1 (PSEN1) and PSEN2 are also risk factors for AD (Ballard et al., 2011).

Diagnosis and treatment

A complete AD diagnosis should include different aspects. An operationalized clinical diagnosis with criteria such as the NINCDS-ADRDA can be employed to distinguish between patients with

AD and people without dementia (DE LEON et al., 2007). However, more specific biomarkers are needed to improve the accuracy for AD. CT or MRI (magnetic resonance imaging) can be used to detect intracranial lesions or disease that may cause AD (Waldemar et al., 2007). A combination of CSF biomarkers such as total tau has improved diagnosis accuracy cognitive testing (Welge et al., 2009). PET with fluorodeoxyglucose measures glucose metabolism and has shown good accuracy (Patwardhan et al., 2004).

Some disease-modifying treatments for Alzheimer's disease have been proposed. Most of them focus on the Aβ protein. A small part of them target tau phosphorylation or tau aggregation (Ballard et al., 2011). However, no therapies have demonstrated enough efficiency through clinical trials up to date.

1.1.2. Charcot-Marie-Tooth Disease

Epidemiology

Charcot-Marie-Tooth disease (CMT) is the most common degenerative disorder of the peripheral nervous system, occurring in 1 out of 2500 people (Dyck and Lambert, 1968). It is also referred as "Hereditary Motor and Sensory Neuropathy" (HMSN).

Most cases of CMT are slowly progressive disorders that usually present in the second decade. Typically, CMT patients display weakness of the foot and lower leg muscles, which may result in foot drop and a high-stepped gait with frequent tripping or fall. Foot deformities are also characteristic due to the weakness of small muscles. With the development of the disease, weakness and atrophy may occur in the hands (Tazir et al., 2014). The severity of symptoms varies in different patients. As the most frequent form of CMT, CMT1A leads to a mild to moderate disability, although some of CMT1A patients have a marked handicap and end up in wheelchairs.

CMT is caused by mutations that cause defects in neuronal proteins. The different subtypes of CMT have various frequencies within distinct populations and can be classified by their clinical, neurophysiological, genetic and pathological features. Its two major subtypes, demyelinating (CMT1) and axonal (CMT2) are usually inherited as an autosomal dominant trait (Harding and

Thomas, 1980a). They can be distinguished by electrophysiological and nerve biopsy studies (Harding and Thomas, 1980b). The duplication of peripheral myelin protein 22 gene (PMP22) on the chromosome 17 is the cause of the most frequent form of CMT1 which is named CMT1A (Lupski et al., 1991). CMTX is an X-linked disorder and sometimes autosomal recessive CMT variants are classified as CMT4 (Yum et al., 2009). At present, more than 75 genes have been shown to be involved in a CMT phenotype.

Diagnosis and treatment

To develop efficient treatment, accurate diagnosis is of importance. Currently, genetic testing is used and recommended by clinicians and relies on nerve conduction velocity assessment, disease inheritance pattern and population frequency (Ekins et al., 2015). CMT can also be diagnosed through symptoms, through measurement of the speed of nerve impulses, through biopsy of the nerve, and through DNA testing. The severity of the disease can be evaluated through composite neurological scores such as Charcot-Marie-Tooth Neuropathy Score (CMTNS) and Overall Neuropathy Limitations Scale (ONLS) which include the tests of impairment, electrophysiology and activity limitations.

In preclinical studies, ascorbic acid was shown to promote myelination in vitro and to decrease *PMP22* expression (Passage et al., 2004)(Kaya et al., 2007)(Schenone et al., 2011), and its mechanism of action in the murine peripheral nervous system has recently started to emerge (Gess et al., 2011). However, no beneficial clinical effects are reported in the clinical trials targeting ascorbic acid. The analyses of PXT3003 in Pharnext (a low dose combination of the three already approved drugs baclofen, naltrexone and D-sorbitol) as a therapeutic candidate are ongoing to evaluate the disease progression in the patients (Attarian et al., 2014).

1.2. Data in Neurodegenerative Disease

Due to the complexity of neurodegenerative diseases, the research progress of the potential disease-modifying therapies is slow. To conduct cost-effective and informative clinical trials, we need multiple types of data to facilitate the diagnosis and evaluation of disease severities (Shaw et al., 2007). Taking the AD research as an example, various data help to identify those individuals

at greatest risk of developing AD, confirming the diagnosis of AD, predictive testing, monitoring disease progression and response to treatment, enriching clinical trials for specific subsets of patients (Weiner et al., 2012)(Bateman et al., 2012). Growing efforts were made to develop and analyse appropriate data in research of neurodegenerative diseases which are linked to the fundamental features of neuropathology and simple to use. Furthermore, on the basis of extensive studies to date, it is likely that a combination of data will provide greater diagnostic accuracy than a single analysis (Weiner et al., 2012). GWAS data are widely utilized to reveal the genetic contributions to the neurodegenerative diseases by the search for associations between quantitative traits in the form of imaging or biomarker data and genetic loci. The volumetric changes to brain led by the neuronal degeneration could be measured by MRI of specific regions such as the hippocampus. The scores of measure scales are also extensively used to evaluate the degeneration of cognitive ability or disability. Different types of data are explained in below.

1.2.1. Genetic data

Genome-Wide Association Studies (GWAS)

The genetic diversity corresponds to the total amount of different genetic features of a species and is also called the gene pool of a species. As a matter of fact, within a species the genomes of all individuals are not identical. Locus is a specified position on the genome and an allele a possible version of the genetic text at a given locus. It is monomorphic when only one allele is possible (i.e. all the individuals share the same genetic text) and polymorphic when there are several possible alleles at the locus. A haplotype corresponds to a set of several alleles located on different loci of the same chromosome. In humans, for a given locus, each parent passes down one allele to the offspring. Each chromosome therefore carries two alleles at a given locus. Genotype is the combination of alleles at a locus. An individual is homozygous at the locus if the two alleles are the same and heterozygous otherwise. Genetic Epidemiology is a science that combines classical Genetics and Epidemiology, which studies the role of genetic factors in determining disease in families and populations. Benefiting from the advance of high-throughput sequencing technology, Genetic Epidemiology has fast developed. In Genetic Epidemiology, a Genome-Wide Association Studies (GWAS) scans the whole genome to find associations between a disease and many common genetic markers. Since the first successful GWAS study published in 2005 (Klein

et al., 2005), GWAS have successfully identified many genetic variants and facilitated the diagnose and treatment of several diseases. GWAS are usually case-control studies. Two groups of unrelated individuals are selected: patients (cases) and non-patients (controls). Their samples of DNA are genotyped using SNP arrays, and after quality-control, genetic markers on the whole genome are investigated. This process is illustrated in **Figure 1**.

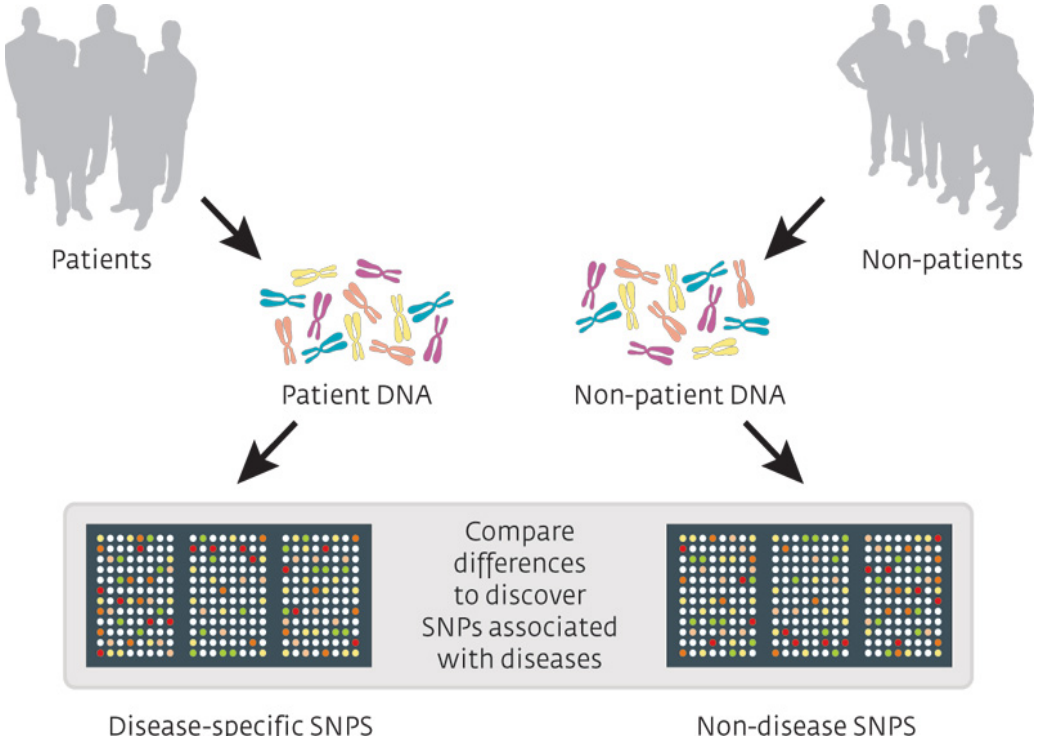

Figure 1: Illustration of the Genome-Wide Association Study (GWAS).

Single Nucleotide Polymorphisms (SNPs)

The most used genetic marker in GWAS is the single-nucleotide polymorphism (SNP). It is the variation of a single base pair of a DNA sequence. Most SNPs involve two possible alleles, which mean two possible versions of the genetic text at the same locus. Until now, in dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/), a free public archive for genetic variation, more than180M of SNPs have been identified. SNPs can arise in a certain population and thus very useful for population differentiate.

GWAS comprises different types of individual information and SNP information. What we would like to test is the association between disease (phenotype) and genotype of SNPs. Usually, there

are two pieces of information. The first contains the SNP annotations: the SNP identifier, the gene(s) and chromosome that they belong to and the position in base-pair. The second corresponds to a table made of the genotypes of all the markers and also includes the information of individual such as the phenotype and the gender. The genotype matrix can be represented as $X = (x_{ij})$ for the n individuals and the p SNPs where $1 \leq i \leq n, 1 \leq j \leq p$.

Each term x_{ij} is the genotype of individual i for the SNP j.

The dosage coding for genotype is under an additive genetic model: x_{ij} = 0, 1 or 2 representing the number of variants alleles. For a SNP with alleles a and A, the code of this SNP can be 0

(genotype aa), 1 (genotype aA or Aa) or 2 (genotype AA). This coding of SNP data is one of the most used in practice. In certain settings, it corresponds to the coding for an additive model, which is usually assumed. The phenotype of the individuals is coded 0 (control) or 1 (case).

For each SNP in GWAS studies, if the allele frequency is significantly altered from control group to case group, this SNP is associated with the disease. Typically, a P-value for the significance of the difference is calculated using test statistics.

Certain features have to be investigated in the quality control process of the SNP data to determine which markers can be reasonably be included in the analysis without leading to incoherent results. The full description and discussion of the quality control process can be found in (Bouaziz, 2012). The main aspects of quality control can include: 1) SNP call rate: the proportion of genotypes per marker with non-missing data. Classically a threshold of 95% is used. 2) Hardy-Weinberg equilibrium: SNPs are summed to follow this equilibrium in a control population. If 3) Minor allele frequency: the minor allele frequency (MAF) of a marker represents the frequency of its less frequent allele in a given population. Typically, a MAF threshold of 1-2% is applied in a lot of GWASs.

GWAS data: The Alzhiemer's Disease Neuroimaging Initiation (ADNI)

In recent years, several Genome-Wide Association Studies (GWAS) were performed to detect genetic loci associated with LOAD (Harold et al., 2009; Potkin et al., 2009; Seshadri et al., 2010).

Alzheimer's Disease Neuroimaging Initiative (ADNI) project is one of them. The goal of the ADNI study is to track the progression of the disease using biomarkers to assess the brain's structure and function over the course of four disease states (Mueller et al., 2005a). Genetic factors play a important role in Alzheimer's disease, and a key aim of this project is providing the opportunity to combine genetics with imaging and clinical data to help investigate mechanisms of the disease. The study population in initial ADANI-1 at baseline is made up of 128 with AD, 415 with MCI, 267 controls and 8 of an uncertain diagnosis. 731 were analyzed using DNA from peripheral blood, and 87 were genotyped using DNA extracted from Lymphoblastoid cell lines. SNP are genotyped with an Illumina Human 610-Quad (= 620901 SNPs). After the quality control, 538830 SNPs satisfying the conditions are kept for further studies. The dataset was also reduced with a minimal loss of information by pruning with Plink (window size = 50 SNPs, shift = of 5 SNPs at each step and threshold correlation coefficient of 0.2). SNPs are considered attached to a gene if they are located within a distance of 20 kb around it. The curated dataset to analyze comprises 16514 genes.

1.2.2. Rating scales

Rating scales are widely used to measure the health outcomes of trials for the treatment of neurodegenerative diseases (Hobart et al., 2007). They are increasingly selected as primary or secondary outcome measures in clinical trials (Mandel et al. 2015; Graham and Hughes 2006) and therefore become the main dependent variables on which decision are made that influence patient care and guide future research. The patient-reported rating scales are included in the US Food and Drug Administration (FDA) scientific requirements in clinical trials (Revicki et al., 2007), which indicates their importance.

Two types of rating scales are commonly used in neurology: single-item scales such as the EDSS (Kurtzke, 1983) and multi-item scales such as the ADAS-cog. The score generated by single item scales is easy to be interpreted but has poor reliability and validity. On the other hand, the multiple item scales where the scores from a set of items are combined to give a single value allow complex variables to be evaluated in parts. Each item in the scales has two or more ordered response categories that are assigned sequential integer scores. Although the scores generated by the multiple items scales are less clinically tangible, the validity and precision of them are

improved because the continuum is divided into more parts. Therefore, the multiple items scales are preferred in clinical trials.

In the neurodegenerative disease process, various domains of abilities become differentially affected, subsequently resulting in progressive functional decline. For example, patients with AD typically perform poorer on tests of memory, language, executive function, and visuospatial ability as part of disease progression (Park et al., 2012). Since most rating scales are unidimensional, only one multiple items scale is not enough to evaluate these declinations comprehensively. In the case of complex disease research such as AD, multiple rating scales are employed to detect the cognitive changes more precisely. In the Alzheimer's Disease Neuroimaging Initiative (ADNI), a study designed to identify biological and clinical markers of AD, 19 scales were included in the neuropsychological battery to measure different abilities, such as Boston Naming Test for memory evaluation and Trail Making Test for executive function evaluation (Mueller et al., 2005a). Many studies in ADNI used only summary scores from brief global scales (Mini-Mental State Examination and Alzheimer's Disease Assessment Scale-Cognitive Subscale). However, their scores do not capture varying levels of change that can occur across different domains. Two rating scales are employed in most of the clinical trials to evaluate the impairment in the CMT disease. The Charcot-Marie-Tooth Neuropathy Score (CMTNS) comprises item measuring impairment such as the sensory symptoms, activity limitations, and electrophysiology. The Overall Neuropathy Limitations Scales (ONLS) is another example in peripheral neuropathies to measure limitations in the everyday activities of the upper limbs and the lower limbs (Graham and Hughes, 2006).

Rating scale data I: the Charcot-Marie-Tooth Neuropathy Score (CMTNS)

A typical feature of CMT1A is the weakness of the foot and lower leg muscles. With regard of this, the Charcot-Marie-Tooth Neuropathy Score (CMTNS) which measures the impairment such as the strength arms and legs was developed and has been used as the primary or main endpoint in most completed clinical trials for CMT1A. The CMTNS is composed of 9 items evaluating different functions related to the disease. The items and the score standards are listed in **Figure 2:**

Parameter	0	1	2	3	4
Sensory symptoms[1]	None	Symptoms below or at ankle bones	Symptoms up to the distal half of the calf	Symptoms up to the proximal half of the calf, including knee	Symptoms above knee (above the top of the patella)
Motor symptoms legs[2]	None	Trips, catches toes, slaps feet. Shoe inserts	Ankle support or stabilization (AFOs). Foot surgery[5]	Walking aids (cane, walker)	Wheelchair
Motor symptoms arms	None	Mild difficulty with buttons	Severe difficulty or unable to do buttons	Unable to cut most foods	Proximal weakness (affect movements involving the elbow and above)
Pinprick sensibility[1,3]	Normal	Decreased below or at ankle bones	Decreased up to the distal half of the calf	Decreased up to the proximal half of the calf, including knee	Decreased above knee (above the top of the patella)
Vibration[4]	Normal	Reduced at great toe	Reduced at ankle	Reduced at knee (tibial tuberosity)	Absent at knee and ankle
Strength legs	Normal	4+,4 or 4- on foot dorsiflexion or plantar flexion	≤ 3 on foot dorsiflexion or ≤ 3 on foot plantar flexion	≤ 3 on foot dorsi and ≤ 3 on plantar flexion	Proximal weakness
Strength arms	Normal	4+,4 or 4- on intrinsic hand muscles[5]	≤ 3 on intrinsic hand muscles[6]	< 5 on wrist extensors	Weak above elbow
Ulnar CMAP (Median)	>6mV (>4mV)	4-5.9mV (2.8-3.9)	2-3.9 mV (1.2-2.7)	0.1-1.9 mV (0.1-1.1)	Absent (Absent)
Radial SAP amplitude, antidromic	≥15µV	10 - 14.9 µV	5 - 9.9 µV	1 - 4.9 µV	< 1 µV

Figure 2: Items and definition of scores in CMTNS**.**

Each component of the CMTNS is scored on a 0-4 point scale, positively correlating with the respective severity of each examined item. The scores of the CMTNS range from 0 (good clinical performance) to 36 (severely affected). Patients are classified according to the scores as mild (CMTNS <=10), moderate (CMTNS 11-20), or severe (CMTNS > 20).

However, the sensitivity of the CMTNS to change and its psychometric properties are still debated. Clinical trials investing the efficacy of ascorbic acid as a therapy confirmed difficulties in measuring the disease worsening over time. A modified version of the scale (CMTNS-v2) was proposed by Murphy in an attempt to reduce the aforementioned effects and to standardize patient assessment (Murphy et al., 2011). Certain items and score standard were modified in this version (**Figure 3**):

Parameter	0	1	2	3	4
Sensory symptoms *	None	Symptoms below or at ankle bones	Symptoms up to the distal half of the calf	Symptoms up to the proximal half of the calf, including knee	Symptoms above knee (above the top of the patella)
Motor symptoms (legs) †	None	Trips, catches toes, slaps feet Shoe inserts	Ankle support or stabilization (AFOs) Foot surgery ‡	Walking aids (cane, walker)	Wheelchair
Motor symptoms (arms)	None	Mild difficulty with buttons	Severe difficulty or unable to do buttons	Unable to cut most foods	Proximal weakness (affect movements involving the elbow and above)
Pinprick sensibility *§	Normal	Decreased below or at ankle bones	Decreased up to the distal half of the calf	Decreased up to the proximal half of the calf, Including knee	Decreased above knee (above the top of the patella)
Vibration/	Normal	Reduced at great toe	Reduced at ankle	Reduced at knee (tibial tuberosity)	Absent at knee and ankle
Strength (legs) ¶	Normal	4+, 4, or 4− on foot dorsiflexion or plantar flexion	≤3 on foot dorsiflexion or ≤3 on foot plantar flexion	≤3 on foot dorsiflexion and ≤3 on plantar flexion	Proximal weakness
Strength (arms) ¶	Normal	4+, 4, or 4− on intrinsic hand muscles **	≤3 on intrinsic Hand muscles **	≤5 on wrist extensors	Weak above elbow
Ulnar CMAP	≥6 mV	4–5.9 mV	2–3.9 mV	0.1–1.9 mV	Absent
(median)	(≥4 mV)	(2.8–3.9)	(1.2–2.7)	(0.1–1.1)	(absent)
Radial SAP amplitude, antidromic testing	≥15 μV	10–14.9 μV	5–9.9 μV	1–4.9 μV	<1 μV

Figure 3: Items and definition of scores in CMTNS-v2**.**

Another modified CMTNS called CMTNS-Mod has also been proposed by adding three functional measures (9-hole peg test, foot dorsiflexion and walk test) and removing four of the initial items ('*Ulnar SNAP*', '*Pin Sensibility*', '*Vibration*' and '*Strength of Arms*') (Mannil et al., 2014).

Rating scale data II: The Alzheimer's Disease Assessment Scale-Cognitive Subscale test (ADAS-cog)

The Alzheimer's Disease Assessment Scale- Cognitive Subscale test (ADAS-cog) is one of the most frequently used tests to measure cognition in AD. It has become the standard primary

outcome measure for evaluating treatments in clinical trials of mild-to-moderate Alzheimer's disease (Ihl et al., 1999). It was designed specifically to evaluate the severity of cognitive and noncognitive behavioral dysfunctions characteristic of persons with Alzheimer's disease. Despite some questioning about its sensitivity (Cano et al., 2010; Ihl et al., 1999), it has proven successful for its intended purpose.

The original 11-item ADAS scale (ADAS-cog 11 or ADAS-classic) was developed by Rosen et al. in 1984 (Rosen et al., 1984). The names and score ranges of the items are listed as in **Figure 4**:

	Score range
Memory and new learning	**0 - 35**
Word recall *(mean number of words not recalled)*	0 - 10
Orientation *(one point for each incorrect response)*	0 - 8
Word recognition *(mean number of incorrect responses)*	0 - 12
Remembering test instructions	0 - 5
Language	**0 - 25**
Commands	0 - 5
Spoken language ability	0 - 5
Naming objects/fingers	0 - 5
Word-finding difficulty	0 - 5
Comprehension	0 - 5
Praxis	**0 - 10**
Constructional praxis	0 - 5
Ideational praxis	0 - 5
Total	**0 - 70**

Figure 4: Items and score ranges in ADAS-cog.

The total score of ADAS-cog classic ranges from 0 to 70 indicating the dysfunction severity increases. These items were designed to assess three cognitive impairment in the memory, language, and praxis cognitive domains (Rosen et al., 1984).

Mohs *et al.* modified ADAS-cog classical to broaden the scope of cognitive domains covered and range of symptoms consistent with mild to moderate AD. This version added the 2 items Delayed Word Recall and Number Cancellation on the ADAS-cog classic for a total of 85 points and is called ADAS-Modified or ADAS-cog 13 (Skinner et al., 2012). The purpose of these additional items was to increase the number of cognitive domains and the range of symptom severity without a substantial increase in the time required for administration. The modified version improved the

responsiveness to the MCI patients. ADAS-cog 13 was included in the neuropsychological battery of ADNI.

1.3. Item Response Theory

Like rating scales data, SNP data from GWAS data also comprise a set of “items”, each having three ordered response categories that are assigned sequential integer scores (0,1 and 2). Typically, the Classical Test Theory is applied to the treatment of items with ordered categorical data, whereby the item scores are summed to give a total score. However, this simple and natural method has its limitations (Hobart et al., 2007). As an important alternative method capable of overcoming the limitations, the Item Response Theory (IRT) does not suppose that each item is equally difficult. Instead, it assumes that the probability of a person achieving a certain score on a test is a consequence of that person's ability on the latent construct and the difficulties of items.

In recent years, IRT has been widely applied in the fields of psychometric, social sciences, education, business and clinical trials. IRT is a statistical framework and provides enriched statistics for categorical data analysis. Therefore, it can also be applied to improve the analysis of data involved in neurodegenerative diseases in different ways. It has been mostly used for evaluation of psychometric properties of the measures in the neuropsychological battery (Burns et al., 2012; Sadjadi et al., 2014). Nevertheless, it can be utilized in the analysis of other types of categorical data such as GWAS data or it be combined with other statistical methods such as information theory. Applying IRT to neurodegenerative diseases data has the potential to help clinicians and researchers to lead to advancements in screening assessments and diagnosis, the measurement of change with disease progression and in response to treatment.

We present below the Item Response Theory and its extensions.

1.3.1. Background

Latent traits

Some variables (e.g. height and weight) in the physical world can be measured directly, where numbers are given to represent quantities of certain properties of some attributes. Other variables (e.g. disability, cognitive function, quality of life) can only be measured indirectly through observable indicators of the attributes that measurements can be made. These variables are often some concepts or notions which need clarification before measurement can take place. They are called “latent traits”.

Psychometrics

The science of measuring the latent traits is referred to as psychometrics. The latent traits concerned by psychometrics are not limited in the psycho-social context. Since the latent traits cannot be measured directly, psychometrics methods collect information on indicator variables associated with the latent trait and numbers are assigned to these variables, to represent the quantities of latent trait.

Measurement level

The goal of all measurement models is to arrange samples on a latent continuum. There are four levels of measurement for assigning numbers to indicator variables described below.

Nominal is that the numbers assigned to objects as labels. For example, in a survey, the genders of respondents are assigned by number: male = 1; female = 2. The nominal numbers are not for comparison.

Ordinal means the numbers assigned to objects indicate their order. For example, the responses to a question: disagree, agree, strongly agree are represented by numbers 1, 2, 3.

Interval is the case when numbers are assigned to object to indicate the amount of an attribute. The numbers on a clock represent interval measurement of time. An absolute zero is not necessary for Interval measurements.

Ratio measurement is the interval measurement with an absolute zero. We can compare not only the distance between the numbers but also the ratios formed by numbers.

These four levels of measurement provide increasing power in the meaningfulness of the numbers.

Rating scale to measure latent traits

To measure a latent trait, the psychometric methodologies are applied to establish a rating scale. Two types of rating scales are commonly used: single item and multiple item scales. In the following, we will focus on multiple item rating scales. Multiple item rating scales comprise a set of items, each of which has two or more ordered response categories that are assigned sequential integer scores. For example, a typical Likert scale provides five response options for each survey item: Strongly Disagree (*SD*), Disagree (*D*), Neutral (*N*), Agree (*A*), and Strongly Agree (*SA*). Each option is coded numerically like $SD = 1$, $D = 2$, $N = 3$, $A = 4$, and $SA = 5$. Multiple item scales are widely used in the fields of psychometric, social sciences, and education.

On an ideal scale for measuring a latent trait, or ability scale, the numbers represent the highest level of measurement, which is the ratio of the latent trait. However, as an absolute zero does not exist for the latent trait, interval measurement could already provide the most information. Also, an ideal scale can link the examinees with the indicators. The latent trait levels of examinees are anchored by the hardest item that they can achieve on the scale, like a real rule. This ability scale can be used to tell how much ability a given person has compared to other persons.

Measurement issues

Some statistical issues of rating scales are concerned by many researches.

Reliability test is to estimate the consistency of the accuracy of measurement. A rating scale is reliable if a measurement consistently gives the same estimate. The way is to determine how much of the variability in the result is due to random errors in measurement and how much is due variability of the true scores, which is the latent trait being measured. In another word, Reliability estimates the quality of the score. Reliability can indicate ways of improving measurement by

minimizing the errors. We have greater confidence in the higher reliable measurements which have a smaller amount of random errors.

Validity is the degree to which a measure is manifestations of the latent variable. A measure does not necessarily result in an accurate reflection of the latent trait of interest. For example, although certain people under anxiety have symptoms like headaches, but headaches may be led by various causes. Therefore, “frequency and duration of headaches” may not be an accurate “representation” of anxiety and it is not a validated scale. Obtaining validity evidence is part of the measurement process.

Invariance means a measure is independent of what it measures. For instance, a meter stick is not affected by the objects it measures, whereas a rubber rope could be stretched to adapt to the object that it measures. Therefore, the rubber band is not an invariant measure. Without invariance, our comparisons across different samples would have limited utility.

Item analysis

Item analysis is an essential domain in psychometrics referring to statistical methods used for selecting items for inclusion in a psychological test. Item analysis provides a way of measuring the quality of items: seeing how appropriate they are for the respondents and how well they measure their ability/trait. Item analysis also provides a way of re-using items in different tests with prior knowledge of how they are going to perform by creating a bank of questions with known properties. The process of item analysis varies depending on the psychometric model adopted; for example, Classical Test Theory or the Item Response Theory will call for different procedures.

The different psychometric models for item analysis and their relations are presented in **Figure 5** :

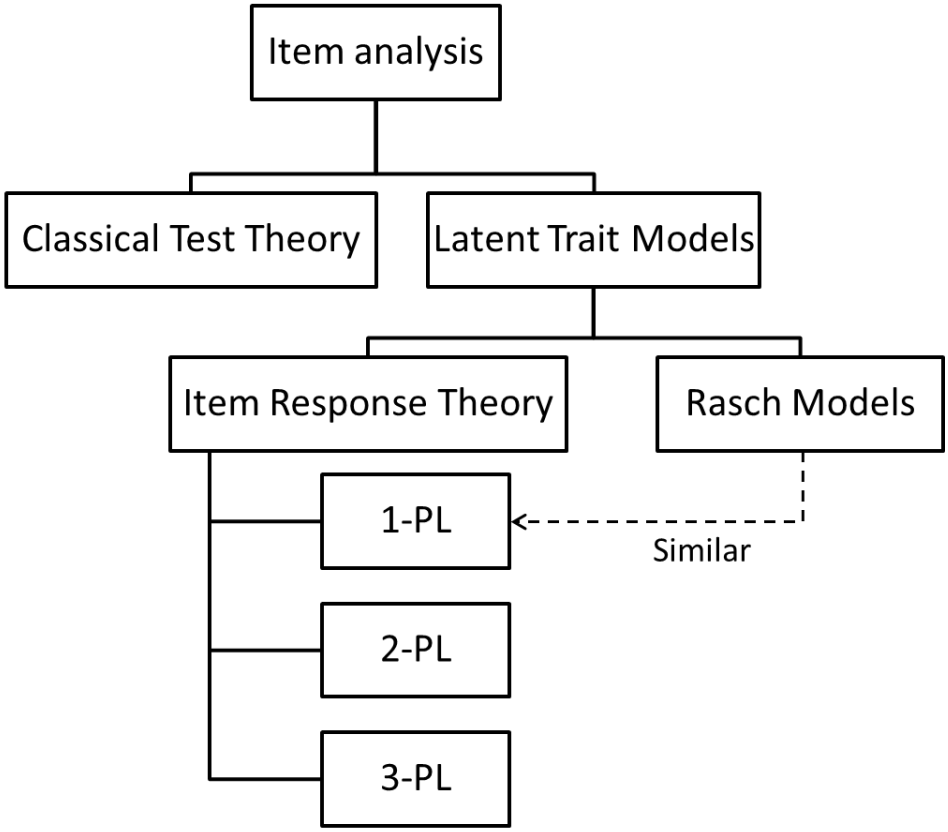

Figure 5: the different psychometric models for item analysis.

In the following discussion, the two main models for establishing a correspondence between our observations and our latent variable will be presented. Classical Test Theory will be first briefly introduced, followed by the explanation of Item Response Theory and its extensions.

1.3.2. Classical Test Theory

Definition

In Psychometrics, the main approache for items analysis has been the Classical Test Theory (CTT), whereby the item scores are summed to give a summed score (also called total or raw score) for each person.

The basic assumption of CTT is that a person has a summed test score and a true score, le latter reflecting a latent trait. The summed score Y_s is an estimate of the true score T_s of this person with some unobservable measurement error e_s

$$Y_s = T_s + e_s$$

The variance of summed score σ_Y^2 is given by

$$\sigma_Y^2 = \sigma_T^2 + \sigma_e^2$$

The reliability coefficient $\rho_{xx'}$ is the ratio of the variance of the observed score and of the true score

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_Y^2} = 1 - \frac{\sigma_e^2}{\sigma_Y^2}$$

Limitations

The advantage of CTT is that it relies on weak assumptions and easy to interpret and calculate. However, this simple and natural approach has two main limitations (Hobart et al., 2007):

The data generated by rating scales, both item scores and summed scores, are at ordinal level. Scoring the items with sequential integers implies same differences at the item level (differences between each response category are implied to be equal) and at the summed score level (a change of one point implies an equal change across the range of the scale, no matter which item is concerned by this change). Consequently, such ordinal scores cannot provide us with a stable frame of reference in terms of the distance between individuals on the ability scale. If there are persons with different ability levels, a less or more difficult test will probably result in different scores.

We do not really know what variables most rating scales are measuring. When the CTT is applied, the latent trait of interest is estimated by a summed score which is difficult to match to each single item to know what an individual can perform. Consequently, individuals with the same summed score may not be able to achieve the same item task. Here is an example given by Bond *et al.* (Bond and Fox, 2013): suppose the items on a phobic anxiety scale include:

1. I am so anxious that I have not left my house for five years;

2. I feel uncomfortable in large crowds, though I do not avoid them.

Presumably the first item represents much higher anxiety than the second; the responses to two items like these may, if verified empirically, more appropriately line up like this:

1. *SD D N A SA*

2. *SD D N A SA*

To establish a reliable rating scale, the information of the relative difficulties of items which is lost in the summed score, must be considered. Therefore, it is difficult to compare results of persons between different tests in the framework of CTT.

1.3.3. Introduction of IRT

The concept of IRT was built around the 1950s by three of the pioneers Frederic M. Lord, Georg Rasch and Paul Lazarsfeld (Hambleton et al., 1991). By opposition to CTT, the Item Response Theory (IRT), which assumes that the probability of success of a person on an item depends on the person's ability and the item parameters, is one of the most important alternative theories in Psychometrics. Sometimes it is referred to as modern psychometrics because in large-scale education assessment and testing programs IRT has become a more popular choice compared to CTT. The person's ability is the estimate of the latent trait of an individual; the item parameters refer to different characteristics of an item. For example, item difficulty is the difficulty level of an item to be achieved by people.

In IRT, people at the same ability level have a certain probability that they will give a correct answer to an item. This probability becomes low for people with low ability and high for those with high ability. The relationship between the probability of a correct response and the person's ability can be described as a monotonic S-shape curve. The item parameters and the person's abilities can be estimated through with the responses of individuals to an item. Therefore, they can be anchored on the same scale and can be compared.

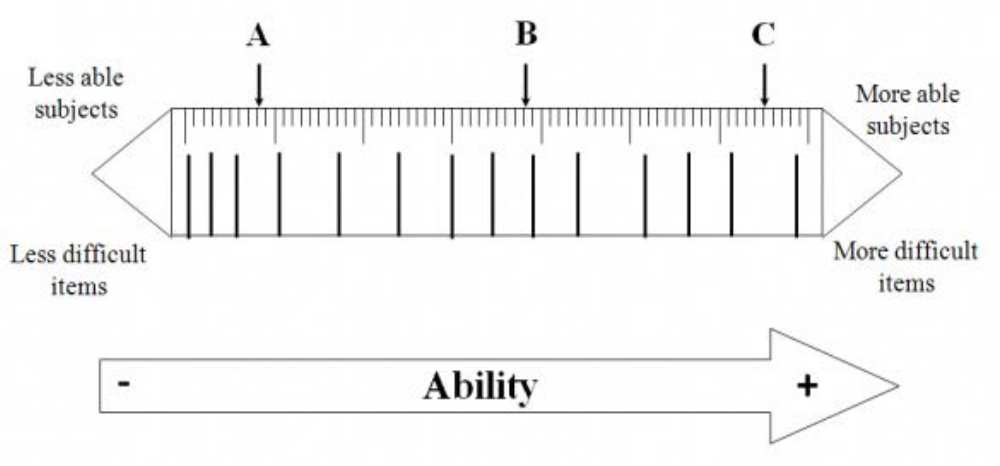

Figure 6: The subjects and items could be located on the same scale in IRT.

Difference between CTT and IRT

Besides providing more sophisticated information about subjects, IRT provides a framework to evaluate the measurement as well as the individual items. It has several advantages. First, it is an interval measurement: transformation preserves the order of raw scores, while the distance between individuals can be made, and not just rank ordering. Second, both the item difficulty and person ability are defined on the same scale. If a person's ability is known, we can predict how that person is likely to perform on an item. The items from different tests can also be placed on the same scale.

Assumptions of IRT

The assumption of models is about the type of data that model applies, and specifies the relationships between observable and unobservable constructs described in the model. Unlike the CTT, IRT model is a falsifiable model, which means it may or may not be appropriate for a particular set of data. In any application of IRT, it is essential to verify if the data satisfy a set of strong assumptions:

Unidimensionality is the most widely used assumption of IRT. It assumes that the response to each item depends on a unique latent trait. This assumption cannot be entirely met because several cognitive, personality and test-taking factors always affect test performance. However, what unidimensionality requires is the presence of one "dominant" component or factor that influences

the measurement. This dominant component is referred to as latent trait. Models assumming more than one latent trait affect the test performance are defined as multidimensional. These models are more complex and have limited application, therefore are not discussed in this context.

Local independence means that, conditionally on the latent person ability, the response of a particular individual to an item depends neither on the responses to other items nor on the responses given by other people to the same item. After taking the abilities of subjects into account, no relationship exists between the responses of subjects to different items. When a scale is local independent, the latent trait is the only factor influencing the responses of a subject. Conditional independence can provide us with statistically independent probabilities for responses of item i and i':

$$P(X_i = 1, X_{i'} = 1) = P(X_i = 1)P(X_{i'} = 1).$$

The third assumption is that the means of the conditional distributions is connected by an S-curve expressing the regression of item score on ability. The curve is referred to as an item characteristic curve or item characteristic function which will be explained in below. There are various mathematic functions in IRT. It is up to the user to choose one of them to serve in the categorical data analysis. In doing so, an assumption is being made that can be verified later by how well the chosen model accounts for the test results.

1.3.4. Item Response Function

In IRT, one of the most important things is to predict the probability of responses. When a dataset satisfies the IRT assumptions, several desirable features are obtained. The estimated ability of a person is not test-dependent; the item parameters are not group-dependent. Therefore, the persons of the same ability have the same probability of giving a correct response to a certain item. This probability can be described as the association between the latent trait level of a person and the probability of a particular item response using a nonlinear monotonic function f:

$$\mathrm{P}(X_{ik}|\theta_i) = f(\theta_i, \beta_k,)$$

where $X_{ik} = x \in \{0, 1, ..., m_k\}$ is an integer random variable for item k indicating the categories and m_k is the maximum score, θ_i corresponds to the ability parameter of person $i \in \{0, 1, ..., n\}$. β_k corresponds the parameters of item k. This function is called the **Item Response Function (IRF)**. IRF is the primary character of IRT theory. Since probability ranges from 0 to 1, a generally logistic function is used, which results in an S-shaped curve. The graphic presentation of IRF is known as the **Item Characteristic Curve (ICC)**. When the ability parameter is high and the difficulty parameter is low, the probability of a correct answer to the item increases.

There are two types of response categories for items in a scale: dichotomous and polytomous.

Dichotomous means there are only two response categories (0 = wrong, 1 = correct) of an item. It is supposed that the probability of success is 0.5 when the ability equals to the item difficulty. If $P(X_{ik} = 1|\theta_i)$ = 0.80, then 80% of individuals with the given θ_i should answer the item correctly. An example of ICC of a dichotomous item is shown in **Figure 7**.

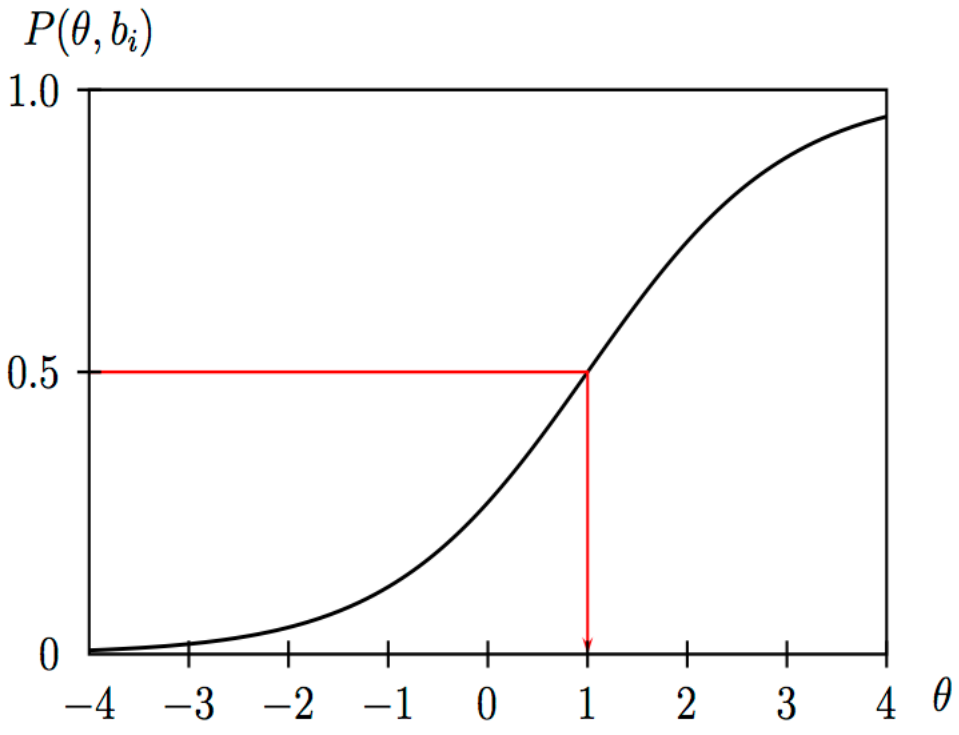

Figure 7: An example of an Item Characteristic Curve (ICC) of a dichotomous item**.**

Polytomous means there are more than two responses categories. Typically they are ordered with increasing level. In the case of a polytomous item, in addition to the ICC, a Category Characteristic Curve (CCC) can be produced for each item, which displays the probability of a person choosing a particular response category based on their level of ability and the difficulty of the item (**Figure 8**). The point between two adjacent categories, where the probabilities of choosing either category are equal, is termed the threshold.

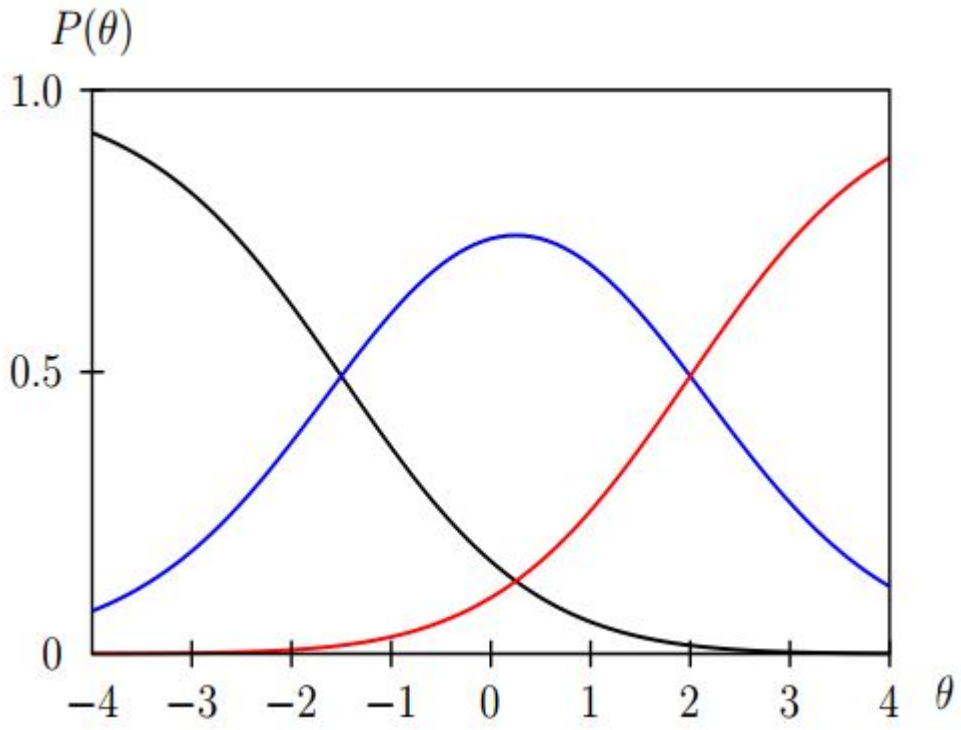

Figure 8: an example of a Category Characteristic Curve of a polytomous item.

1.3.5. IRT Model parameters

A primary distinction among the most popular unidimensional IRT models is the number of parameters used to describe items. The choice of model is up to the user, but this choice can be verified later by examining how well the model fits the data. There are three main types of IRT models with one-, two-, and three-parameters and they are briefly described as following. For simplicity, we focus on the dichotomous model here as an example.

One-Parameter Model

The One-Parameter Model (1-PL) assumes that all items relate the latent trait equally and items vary only in difficulty. In the dichotomous model, the probability that person i gets a correct answer to item k is described as:

$$P(X_{ik} = 1|\theta_i) = e^{D(\theta_i - \beta_k)}/(1 + e^{D(\theta_i - \beta_k)})$$

where $X_{ik} = x \in \{0, 1\}$. D is a scaling factor. Typically, it can be set as 1.7 to make the logistic function essentially the same as the normal ogive model. $\theta_i$indicating that the person abilities are modelled as a sample from a normal distribution for estimating the item parameters. After the item parameters have been estimated, the abilities of individual person are estimated for reporting purposes.

The item difficulty parameter β_k represents the item k location on the same logit scale as the latent trait. It is the location where the ICC has its maximum slope, and where the value is half-way between the asymptotic minimum $\mathrm{P}(-\infty)$ and asymptotic maximum $\mathrm{P}(\infty)$ of the ICC. In the case of 1-PL, $\mathrm{P}(-\infty)$ of ICC equals to 0 and $\mathrm{P}(\infty)$ equals to 1, thus $P(\beta)$ equals to $\frac{\mathrm{P}(-\infty)+\mathrm{P}(\infty)}{2} = 0.5$, means that β equals to the latent trait level θ needed to have 50% chance of endorsing an item. Item difficulty parameters determine the location of the ICC curve. The higher the item difficulty, the higher on the latent trait level that a person needs to be in order to endorse the item, and the ICC curve of this item would be close to right end on the logit scale.

Two-Parameter Model

The Two-Parameter Model (2-PL) IRT model extends the 1-PL model by estimating an item discrimination parameter α:

$$P(X_{ik} = 1|\theta_i) = e^{D\alpha_k(\theta_i - \beta_k)}/(1 + e^{D\alpha_k(\theta_i - \beta_k)})$$

This discrimination parameter is similar to a correlation between the item and total score. It ranges typically from 0.5 to 2. In 2-PL, items vary both in their discrimination and difficulty level. The discrimination parameter illustrates the capacity of an item to discriminate between contiguous trait levels near the inflection point. Therefore, more discriminating items provide greater information about the latent trait than do less discriminating items. Discrimination

parameter decides the maximum slope of the ICC curve. Items with high discriminations are better at differentiating persons around the latent trait location point; smaller changes in the latent trait lead to substantial changes in probability, and the ICC curve is sharper.

Three-Parameter Model

The Three-Parameter Model (3-PL) model extends the 2-PL model by including a pseudo-guessing parameter c, which adjusts for the impact of chance on observed scores:

$$P(X_{ik} = 1|\theta_i) = c + (1 - c)\, e^{D\alpha_k(\theta_i - \beta_k)}/(1 + e^{D\alpha_k(\theta_i - \beta_k)})$$

This model assumes that a person with very low latent trait level may still have a small probability of choosing the correct answer by guessing, raising the lower asymptotic minimum $P(-\infty)$ Of the function to c. This model is mostly used with multiple choice testing where guessing could be a factor in test performance. In 3-PL, the probability of the response at $\theta = \beta = (1 + c)/2$.

An ICC curve in a 3-PL IRT model could be represented as in **Figure 9**:

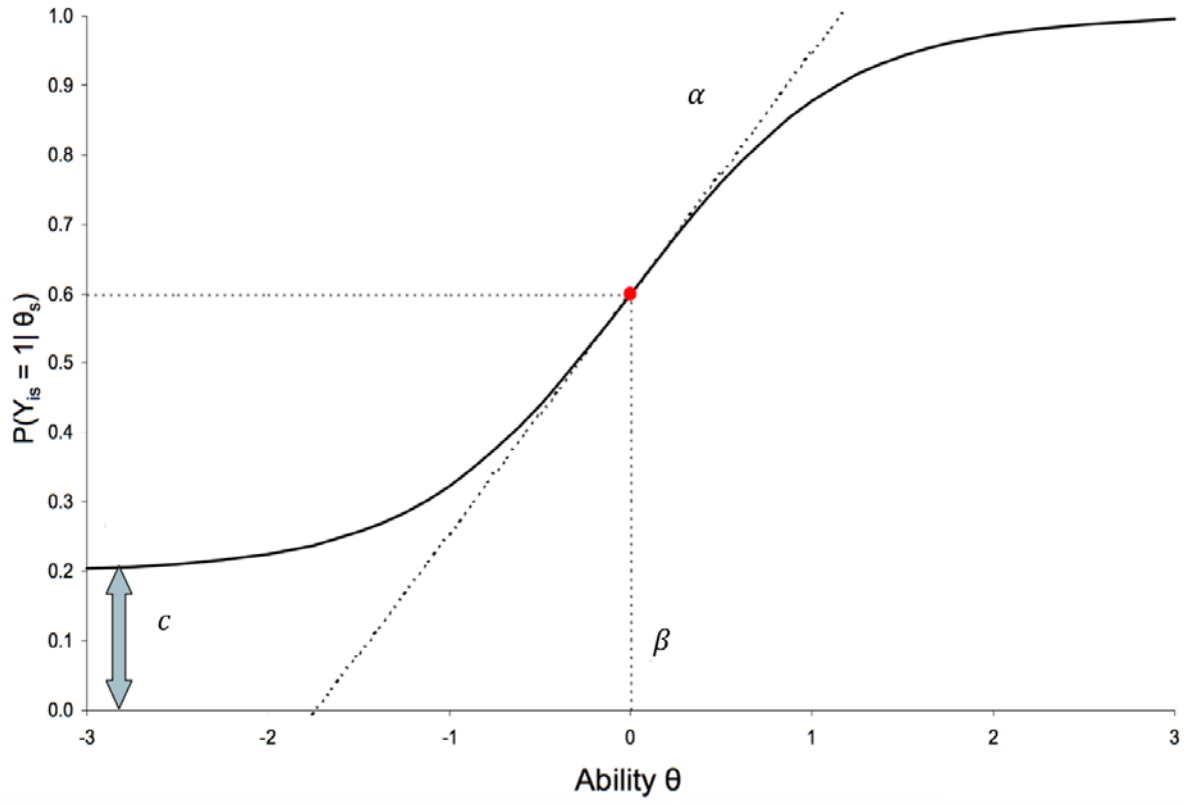

Figure 9: Illustration of the different item parameters in a 3-PL model.

From this figure, it is seen that for a 3-PL dichotomous ICC, the item difficulty β is the location where the ICC has its maximum slope; item discrimination parameter α is the maximum slope; pseudo-guessing parameter c decides the lower asymptote of the curve.

1.3.6. IRT Model extension

For polytomous items having more than two categories, several models based on dichotomous models are developed. Instead of being names with the number of item parameters, polytomous models get called different names. The main difference between these models is how they use multiple thresholds per item.

1-PL extension

Under the 1-PL IRT model which assumes the equal discrimination across items, two polytomous models are developed: Rating Scale Model (RSM) and Partial Credit Model (PCM).

Rating Scale Model (RSM)

RSM is developed by Andrich in 1978 (Andrich, 1978). Under RSM, all items share the same scale structure, which means 1) the items have the same number of thresholds and 2) the difference between any given threshold location is equal across items. Each item is assumed to contain the same amount of information and have the same slope. In RSM, each item is described by the latent trait level, its difficulty, and thresholds that identify boundaries between the ordered categories.

The probability of the person i to endorse the lth categories for the item k is given by

$$P(X_{ik} = l|\theta_i) = \frac{exp\sum_{x=0}^{l}(\theta_i - (\beta_k - \delta_k))}{\sum_{j=0}^{m_k} exp\sum_{k=0}^{j}(\theta_i - (\beta_k - \delta_k))}$$

Where θ_i is the person parameter on the latent trait scale, β_k is the difficulty of item k and δ_k is the kth threshold location of the rating scale, m_k is the maximum score and is identical for all the items.

Partial Credit Model (PCM)

PCM is developed by Masters in 1982 (Masters, 1982). Compared to RSM, PCM assumes that each item has a unique scale structure. Items could have different thresholds and numbers of categories. PCM models the probability of adjacent response categories directly. δ is the threshold of two adjacent categories where the next category becomes more likely – not necessarily 50%. For example, for an item k with 4 categories, 3 thresholds dividing the item into a series of binary items without order constraints beyond adjacent categories: $\delta_{k1}(0\ vs.1)$, $\delta_{k2}(1\ vs.2)\ and\ \delta_{k3}(2\ vs.3)$.

The probability of the person i to endorse the lth categories for the item k is given by

$$P(X_{ik} = l|\theta_i) = \frac{exp\sum_{x=0}^{l}(\theta_i - \delta_{kl})}{\sum_{j=0}^{m_k} exp\sum_{l=0}^{j}(\theta_i - \delta_{kl})}$$

Where δ_{ki} is the kth threshold location of the item i, m_i is the maximum score for item i. The value of δ_{k0} is chosen for computational convenience that is: $\sum_{l=0}^{0}(\theta_i - \delta_{kl}) \equiv 0.$

The different of the thresholds between RSM and PCM is illustrated in **Figure 10**.

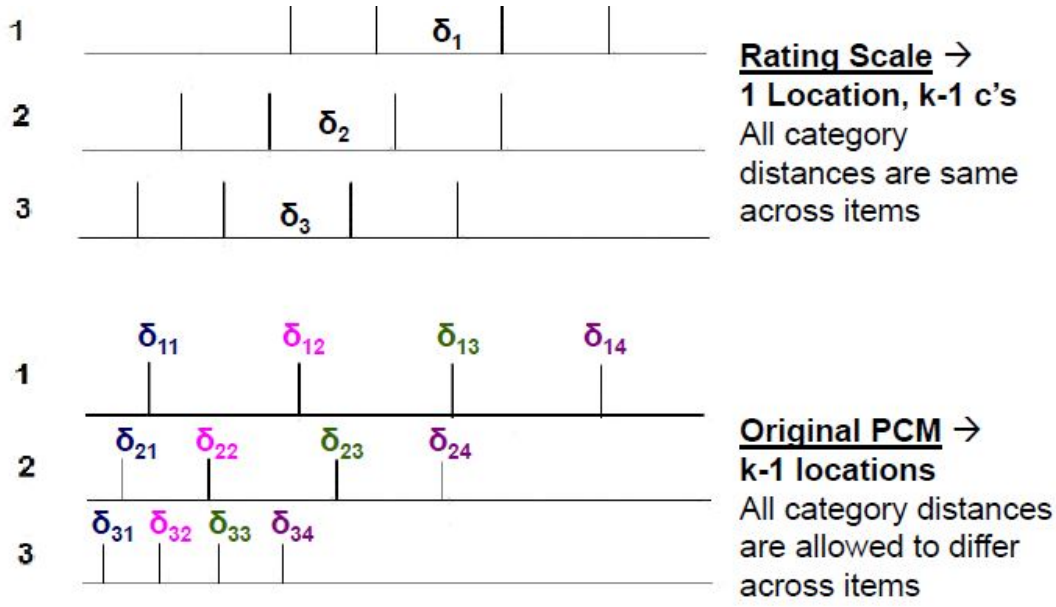

Figure 10: The thresholds of items with five categories in model RSM and PCM.

2-PL extension

Graded Response Model (GRM)

For polytomous data in which the item responses are characterized into ordered categories, Samejima (Samejima, 1970) introduced the Graded Response Model (GRM), an extension of the 2-PL IRT model. It is ideal for items with clear underlying response continuum. In the GRM, items need not have the same number of response categories. It is a cumulative logit model in which the probability of each response is computed by the difference between models of categories. For instance, for an item k with 4 categories, there are 3 thresholds of categories dividing the item into a series of binary items: $\beta_{k1}(0\ vs.\ 1,2,3)$, $\beta_{k2}(0,1\ vs.\ 2,3)\ and\ \beta_{k3}(0,1,2\ vs.\ 3)$.

In GRM, for item k, the probability that the person i endorses the lth or higher response categories is given by

$$P(X_{ik} \geq l|\theta_i) = \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{kl})}}$$

Where the X_{ik} is the ordinal manifest variable with L_k possible response categories. The threshold β_{kl} can be considered as the difficulty of responding with category l or higher for item k with $\beta_{k1} <\ldots< \beta_{kl} < \beta_{k,L_k-1}$ and $\beta_{k,L_k} = \infty$. α_k is the discrimination parameter of item k. θ_i is the person parameter. Therefore, the probability of observing lth categories is

$$P(X_{ik} = l|\theta_i) = P(X_{ik} \geq l|\theta_i) - P(X_{ik} \geq l + 1|\theta_i)$$

$$= \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{kl})}} - \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{k,l+1})}}$$

where we take $P(X_{ik} \geq 0|\theta_i) = 1$.

In the case of 4 categories, $P(X_{ik} = 0) = 1 - P(X_{ik} = 1)$, $P(X_{ik} = 1) = P(X_{ik} = 1) - P(X_{ik} = 2)$, $(X_{ik} = 2) = P(X_{ik} = 2) - P(X_{ik} = 3), (X_{ik} = 3) = P(X_{ik} = 3) - 0$.

Compared to PCM, GRM will force the categories threshold parameters to be ordered.

1.3.7. Parameters estimation

The purpose of IRT is to estimate the latent trait on the person who takes the test, as well as the properties of the items. Therefore, the estimation of the person's abilities and item parameters is the most important task of IRT.

Person parameter Estimation

Take the model of dichotomous item as example, the probability of a response vector X is given by

$$P(X|\theta_i) = \prod_{k=1}^{K} {P_k}^{x_k}(1 - P_k)^{(1-x_k)}$$

Where P_k is short for $P(X_{ik} = l|\theta_i)$ and x_k is the binary response to item k.

Maximum likelihood is utilized in the IRT parameter estimation. Once the response X_i of person i is observed, this expression becomes a likelihood function

$$L(X_i|\theta_i) = \prod_{k=1}^{K} {P_{ik}}^{x_{ik}}(1 - P_{ik})^{(1-x_{ik})}$$

The estimation is based on the assumption of local independence. Item responses are independent after controlling the latent trait level, which means the joint probability (likelihood) of items in a test is just the multiplication of the probabilities of each item. Therefore, the log likelihood function can be written as

$$lnL(X_i|\theta_i) = \sum_{k=1}^{K} (x_{ik} lnP_{ik}) + (1 - x_{ik})\ln(1 - P_{ik})$$

In the following, we use lnL to represent the log likelihood function for short.

The value of θ that makes the lnL for an individual a maximum is defined as the maximum likelihood estimate of θ_i for that individual. This equation cannot be solved directly and the most popular approximation methods is the Newton-Raphson procedure. It happens that the ability estimation procedure fails to converge when a person answers either all items in a test correctly or incorrectly.

Item parameter estimation

When the person parameters θ are known, the estimation of item parameters is straightforward and similar to the procedure of person parameter estimation. The difference is the lnL for an item is multidimensional for the item parameters. In the case of the 3-PL model, the item parameters include item difficulty, item discrimination and pseudo-guessing parameters. The values that correspond to the maximum value of a surface in three dimensions must be found. It could be done with the multivariate form of Newton-Raphson procedure by finding the first derivative of the likelihood function with respect to each of the parameters, setting their derivatives to 0 and solving simultaneously the nonlinear equations.

For 1-PL IRT model, since the raw scores are the sufficient statistic, which means the estimation can be done without requiring any further data, we can estimate the parameters by conditional

maximum likelihood. For 2-PL and 3-PL, the joint maximum likelihood or the marginal maximum likelihood could by applied. The joint maximum likelihood estimation is currently the most widely used. The marginal or Bayesian estimation have the potential for solving some of the problems encountered with the joint maximum likelihood procedure.

1.3.8. Information function

Maximum likelihood estimators have several properties of importance. In general, as the sample size and number of item increase, the estimator converges to the true values and asymptotically normally distributed.

Test information function

In IRT, the precision of the ability parameters θ estimate is of concerned. In this context, the precision is measured by the variability of the estimates around the value of the parameter. The person parameter estimator $\hat{\theta}$ is asymptotically normal distributed with mean θ and variance $\mathrm{I}(\theta)^{-1}$ where $\mathrm{I}(\theta)$ is the **test information function** given by

$$I(\theta) = -E\left[\frac{\partial^2 lnL}{\partial \theta^2}\right] = \sum_i I_i(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

Where the $P_i(\theta)$ is the IRF of the item i under the ability level of θ, $Q_i(\theta) = 1 - P_i(\theta)$. P_i' is the derivative of the $P_i(\theta)$. In IRT, the reliability of a measurement is evaluated by item information function.

Therefore, the standard error of θ estimation equals to

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\mathrm{I}(\theta)}}$$

If the $SE(\hat{\theta})$ is large, the person's ability is not estimated precisely enough. In the 2-PL model, the discrimination parameter would be correspondingly small.

Item information function

The amount of information is influenced by the quality and number of items in a scale. The contribution of each item to the total information is addictive. The smaller the items variance, the greater the information. The **item information function** for item i is given as

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

The item information functions are generally bell shaped as shown in **Figure 11**. It depends on the slope of the IRF: the greater the slope and smaller the variance. The item information function arrives at its maximum value at the point where the ICC has the maximum slope, which means the probability to have a correct response equals to 0.5. It can be deduced that the maximum value of the item information is 0.25. The curve of the item information decreases when the ability becomes either smaller or greater than the item difficulty. With this feature of item information, we can see the item provides most information for which ability level.

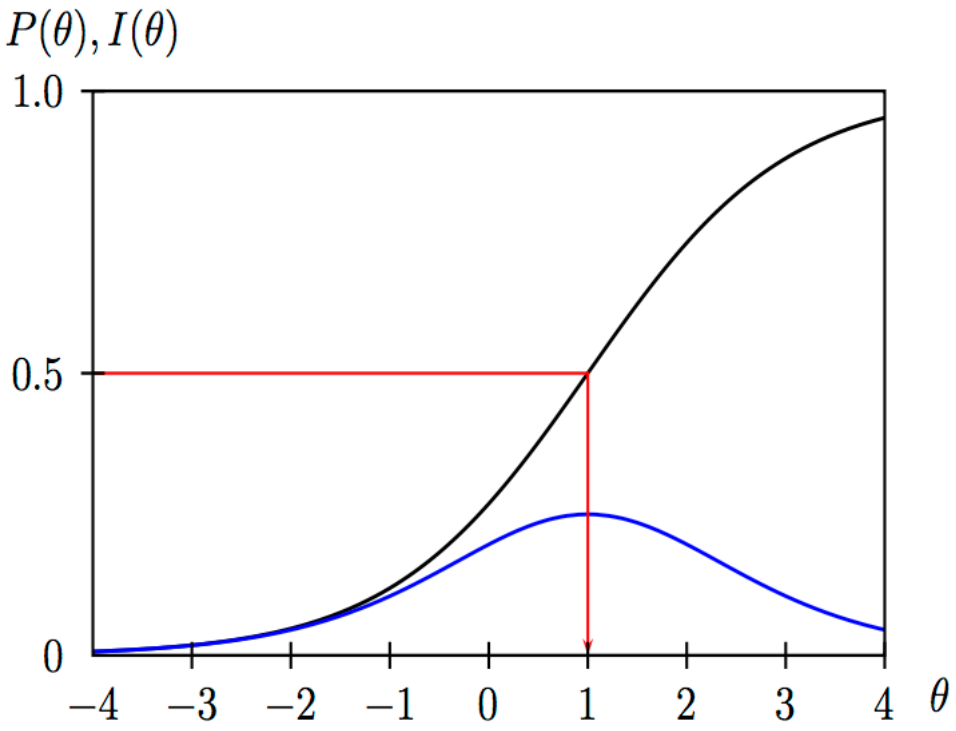

Figure 11: Item Characteristic Curve (in black) and item information function (in blue) of the 1-PL.

The test information is more useful than item information, as the latter is relatively small. The test information curve could be used to diagnose the construction of scale.

Test information curve peaks at some points on the ability scale where the test measures the certain ability level most precisely (**Figure 12**). The shape of the desired test information curve depends upon the purpose for which a test is designed. A test would be best for the people whose ability fall around the peak of the curve. Generally, a test information curve is peaked at the moderate ability level and decrease when the ability is low or high. When the test information curve is rather flat, this test estimates the ability with nearly equal precision.

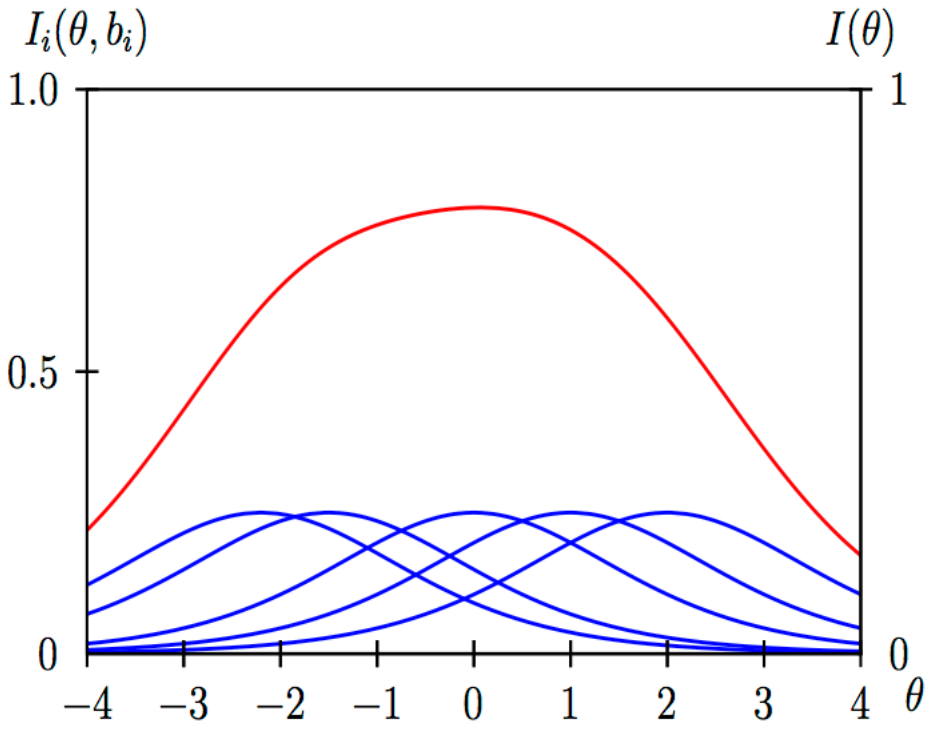

Figure 12: Test Characteristic Curve (in black) and item information function (in blue) of the 1-PL.

1.3.9. Implementation

Several programs are available for IRT analysis. MULTILOG (Reise and Yu, 1990) and PARSCALE (French and Dodd, 1999). A third program, WINSTEPS (Linacre, 2006), is also

noted briefly. There are some differences between these programs. For polytomous data, MULTILOG requires that the number of response categories across items remains the same in a test, whereas the PARSCALE and WINSTEPS allow different numbers of response categories in a test. The WINSTEPS provides figures such as the ICCs.

However, these programs are generally not user-friendly and not easy to be implemented. An R package *ltm* (Rizopoulos, 2006) is available for the analysis of dichotomous data and polytomous data using IRT. It includes the 1-PL, 2-PL, 3-PL and the GRM.

1.4. Rasch Model

The Rasch model is established by a Danish mathematician Georg Rasch in 1960s (Rasch, 1960). It is considered as a special case of the IRT. Mathematically, Rasch models are similar to the most basic IRT model (1-PL): in the basic Rasch model for dichotomous data, the probability for person i get a correct answer for item k is described as:

$$P(X_{ik} = 1|\theta_i) = e^{\theta_i - \beta_k}/(1 + e^{\theta_i - \beta_k})$$

where $X_{ik} = x \in \{0, 1\}$.

For polytomous data, the 1-PL IRT models extensions such as the RSM and PCM can be applied as Rasch model.

1.4.1. Assumptions

Rasch model shares the same assumptions of IRT, which are unidimensionality, local independence and the nature of item characteristic curve. Furthermore, Rasch model has one more assumption which is not shared with other IRT models.

Invariance means that item difficulties remain the same across different groups, such as age or gender. Since the probability of an individual selecting a correct answer to an item depends only on the ability and the item characteristic parameters, this probability is independent of the distribution of individual ability in the population of interest.

Given the dichotomous case of Rasch model, it is easy to demonstrate that

$$\log\left(\frac{P}{1-P}\right) = \theta_i - \beta_k$$

Where $P = P(X_{ik} = 1|\theta_i)$

It shows that the distance between the person's ability and the item difficulty is expressed as the log odds ratios of the probability of getting a correct response for the person. For the same item, the difference between the log odds for two persons would be

$$\log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_2}{1-P_2}\right) = \theta_1 - \beta - (\theta_2 - \beta) = \theta_1 - \theta_2$$

This indicates that the difference is item-free. Similarly, it can be demonstrated that the difference between the log odds of two items under the same ability level is person-free. This property allows us to anchor the item and person on the same ability scale which uses the log odds unit or logit. This logit scale is sometimes also referred to "latent continuum". This scale has a zero midpoint and spreads to positive and negative infinities.

1.4.2. Difference between IRT and Rasch model

However, there are some important differences in interpretation of the result. As frameworks of analyzing measurement data, IRT emphasizes the primacy that the model fit the observed data, while in the Rasch model is superior: data which does not fit the model is discarded (Andrich, 2004). IRT methods include additional model parameters such as the discrimination parameter to reflect the pattern of the observed data, whereas the Rasch model only estimates the latent trait under the condition that both the whole data and the person/item fit the model. Rasch model provides tools to diagnose the measurement and responses, misfitting items or persons may be excluded from the data set. To summary, IRT can be seen as an exploratory approach, on the contrary of the Rasch model which is more a confirmatory analysis.

In IRT, persons are incidental parameters. The latent trait distribution of person sample is conceptualized as normally-distributed $N(0,1)$. $P(X_i)$ is the overall probability of success by

person distribution on item i. Therefore, the estimation of the parameters is computationally simpler. In Rasch model, the latent trait of each person is parameterized individually. $P(X_i)$ is the probability that certain person to have the trait level to be correct on an item i. Also, in Rasch model, there is no scaling parameter D. The ICC is modelled to be parallel with a slope of 1.

1.4.3. Parameters estimation

The main idea of the Rasch model is that the raw score is sufficient statistic for item estimation, and the item total is sufficient for person estimation. Like in the IRT, parameters are estimated by maximum likelihood. Several maximum likelihood estimation methods could be applied for item difficulties estimation, such as conditional maximum likelihood and marginal maximum likelihood. Meanwhile, person parameters are estimated by maximum likelihood. Other than assuming the normal distribution of person ability in the IRT, Rasch model assumes the independence of latent trait level of everyone. Item difficulties and person parameters are estimated interactively.

1.4.4. Rasch analysis

Conducting a Rasch analysis means analysing response data using the Rasch model. The analysis process produces a range of diagnostic information that can be used to determine how well each item contributes to the measurement of the latent trait and in doing so helps inform regarding scale validity and its possible improvement. With tests of fit between the data and the model, Rasch analysis can tell if it is justified to take the total score for person ability evaluation. If the data satisfies a set of requirement, including the basic assumptions and fit of items and persons, the total score could be used for the estimation of person abilities and item difficulties. These parameters, which are more readily than the raw score, can then be used for comparison or analysis. In the practical, data never fit perfectly the model. When the data does not fit well the model, it is important to be able to diagnose where the misfit is the worst. In this case, the use of total score for parameter calibration on the latent trait continuum should be considered carefully.

Therefore, the essential of Rasch analysis is to be in dynamic and interactive control of an analysis and to be able to see if the responses could be valid from the evidence.

Rasch analysis provides an integrated framework of tests. One single statistic is not generally to decide whether a set of data fit the model. The three basic assumptions of the Rasch Model, that of local independence, unidimensionality and invariance should be examined first. Once the assumptions are met, it is possible to use the Rasch model to further evaluate the scale by investigating overall goodness-of-fit, reliability, the fitness of individuals or items, and consistency of items.

Assumption validation

Unidimensionality can be assessed by creating two subsets of items using a Principal Component Analysis (PCA) of the item residuals, with those loading negatively forming one set, and those loading positively forming the second set. Each person parameter estimated from one set of items is then compared to those derived from the other set of items using a *t*-test. If less than 5% of these tests are significant at the 5% level, then unidimensionality is supported (Smith, 2002). Another approach to examining unidimensionality is to apply a generalization of the Martin-Lof test to the two subsets of items defined previously (Christensen et al., 2002). A non-significant *p*-value for this test at the 5% level supports the assumption of unidimensionality.

Local independence is examined by the residual correlations between items, which should be no more than 0.3 for each pair of items (Andrich, 2010).

Invariance is assessed through an analysis of variance of the residuals where the key group of interest is the main factor. If the inter-person-group variance is statistically significant, the item bias is called Differential Item Functioning (Burns et al., 2012; Hanson, 1998). When it is present, the probability of an item response cannot be explained totally by the person and item parameters.

Fitness of test

Once the three assumptions of local independence, unidimensionality and invariance are met, it is possible to use the Rasch model to further evaluate the scale by investigating overall

goodness-of-fit, reliability, the fitness of individuals or items, and consistency of items, as introduced following:

Overall goodness-of-fit. The Andersen's likelihood-ratio test (Andersen, 1973) shows high power and acceptable type-I error rate in Rasch Model estimation (Suárez-Falcón and Glas, 2003). To perform this test, subjects are split into $g = 1, \ldots, G$ score-level subgroups in which a conditional likelihood is computed and compared to the total conditional likelihood computed in the complete sample of subjects. A non-significant *p*-value for this test indicates goodness-of-fit for the Rasch model.

Reliability of the CMTNS scale is estimated by the Person Separation Index (PSI) given by the proportion of true variance relative to the true and error variance. In practice, it measures the internal consistency and the discrimination power of the scale, i.e. the ability of the scale to discriminate amongst persons with different levels of the trait. It is equivalent to the Cronbach's alpha (Cronbach, 1951), but it uses the person estimates in logits instead of the raw scores. A PSI value greater than 0.7 is considered as acceptable.

Item fit can be assessed by several indicators. The residual item fit statistics are expected to approximate a Normal distribution (mean close to 0 with an SD close to 1), which is tested using a chi-square test (Kersten et al., 2014). A significant chi-square test based *p*-value may indicate misfit. In parallel, a similar analysis could be performed for the test of person fit. Then, fit statistics can be computed and focus on two aspects: infit (means inlier-sensitive fit) and outfit (means outlier-sensitive fit). Infit is more sensitive to the overall pattern and less influenced by outliers, and thus infit problems are more of a threat to measurement than outfit ones. Infits and outfits are reported in both mean squares and standardized fit *t*-statistics. The mean squares indicate the amount of distortion of the measurement system whereas the *t*-statistics indicate how likely the item is misfit (Masters, 1982). Mean-squares greater than 1.3 indicates underfit to the Rasch model; mean-squares less than 0.7 indicate overfit to the Rasch model. High *t*-statistics (> 2.0) show that the item distorts or degrades the measurement system as underfit while low *t*-statistics (< -2.0) mean data are too predictable or overfit, but not degrading. Underfit and overfit to the model have different implications for measurement. Underfit degrades the quality of the measurement and should prompt reflection on its cause. Overfit might mislead one into

concluding that the quality of the measure is better than it really is, and has less practical implication than underfit (Green and Frantom, 2002).

The consistency of items. A particularly useful output of the Rasch analysis is the person-item map (also sometimes referred to as `Wright map'). This map displays the difficulty of the items on the same latent dimension as the impairment of the patients. For each item, a threshold of a category is defined as the location at which the cumulative probability of selecting this category versus all the other options reaches 0.5. In doing so, thresholds should follow the same order as categories. A disorder of categories in an item occurs when the ordinal numbering of categories is not in accord with their fundamental meaning or when individuals have difficulties in consistently discriminating categories. In this case, the disordered categories should be rearranged and Item Characteristic Curves representing the probability of selecting each category for one item can be plotted to examine whether this disorder item from under or over-selection of one category.

1.4.5. Implementation

Several programs of Rasch model analysis are available. Winsteps (Linacre, 2006) and RUMM are the main paid software used in the research. They can all be utilized for the dichotomous and polytomous Rasch model analysis.

eRm is an R package for Rasch model and analysis (Mair and Hatzinger, 2007). It provides multiple Rasch models estimation, including RSM and PCM. Also, it contains a simulation module to generate response data matrix for different Rasch scenarios. Furthermore, it provides variant choices for the statistical tests of the parameters fit analysis and global fit analysis.

1.5. Objective and Outline

In this thesis, we applied Item Response Theory and its extensions in different ways to analyze the GWAS data and rating scales data in neurodegenerative diseases to ameliorate the severity evaluation and better understand mechanisms of these diseases. This manuscript is divided into five chapters. After this first introductory chapter, the three chapters present three principal works of this thesis. Each chapter begins with an introduction which introduces the context and

summarizes the methods and results of the article following. Here is a brief overview of each chapter.

The second chapter recounts the application of the Rasch analysis on the CMTNS. Rasch analysis is a typical utilization of the modern psychometric theory on the rating scale. We detail an analysis pipeline using the Rasch model to evaluate the psychometrical properties of the CMTNS in a French cohort of CMT1A patients. After verification, if CMTNS satisfies the basic assumptions of IRT, we evaluates the validity of items on the assessment of the disease severity.

In the third chapter, the Rasch model has been applied to a new type of data: GWAS data as a part of a multi-marker genetic association test. This method summarizes the categorical genotypes of SNPs by Rasch model into a genetic score that can be used for association analysis. Different sets of simulations were carried out to compare the Rasch model based association test with other existing methods. Then this method was applied to a GWAS dataset of Alzheimer's disease to explore disease associated genes.

The fourth chapter describes a novel method to estimate the mutual information of items in a rating scale using the IRT model. The purpose of this study is to select sensitive items in the ADAS-cog, that is the most used cognitive functioning measures in Alzheimer's disease. Using the ADAS-cog scores in the ADNI study, we estimated the mutual information of each item as an indicator of sensitivity and compared it with other IRT-based item statistics for the item evaluation and the rate of change of the composite scores chosen by different methods.

Finally, the fifth chapter is a general discussion. After a summary of the main results of this thesis, I will discuss the possible perspectives of the future works.

Chapter 2

Rasch Analysis on Rating Scale

2.1 Introduction

Charcot-Marie-Tooth disease is the most common inherited disorder of the peripheral nervous system without approved treatment. As a main efficacy endpoint for clinical trials of Charcot-Marie-Tooth disease, the scale Charcot-Marie-Tooth Neuropathy Score (CMTNS) is questioned for its sensitivity to change and its psychometric properties are still under debate. To improve the research on efficacious disease-modifying treatment, a clinically meaningful efficacy endpoint is crucial. One well-accepted way to provide such evidence of validation is to perform a Rasch model analysis, which has been widely employed in clinical scale construction and validation. In the first part of this thesis, I evaluated the psychometrical properties of the CMTNS using Rasch analysis. The purpose of this study is to validate the CMTNS on Charcot-Marie-Tooth 1A patients and propose a possible modification.

This result was published on *Plos One* with the title "A Rasch Analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a Cohort of Charcot-Marie-Tooth Type 1A Patients".

2.2 Methods and Results

We used the Rasch analysis to evaluate the CMTNS scale with a French cohort of 277 CMT1A patients. The Rasch analysis provides an integrated framework for scale evaluation. When a scale could satisfy the assumptions of Rasch model, Rasch analysis was able to provide a range of diagnostic information of this scale. First we tested the three basic assumptions: unidimensionality, invariance and local independence of the scores of CMTNS. Once these

assumptions were met, we could investigate overall goodness-of-fit, reliability, the fitness of individuals or items, and consistency of items in this scale. The polytomous version of Rasch model: Partial Credit Model was applied to the study.

Through the analysis, we found the CMTNS a valid measurement for CMT1A: the three main assumptions of the Rasch were met; the scale showed good overall fit to the Rasch model and an acceptable reliability. Most of items had good fitness to the model, except that two items showed overfit and 3 items had disordered categories. As a limitation, our results pointed out that the items of CMTNS were more suitable for assessing moderate to severe forms of the disease. Therefore, further refinement of the CMTNS such as adding items and/or categories for mild-to-moderate severity assessment is certainly worth consideration.

2.3 Manuscript

PLOS ONE

RESEARCH ARTICLE

A Rasch Analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a Cohort of Charcot-Marie-Tooth Type 1A Patients

Wenjia Wang[1,2]*, Mickaël Guedj[2], Viviane Bertrand[2], Julie Foucquier[2], Elisabeth Jouve[3], Daniel Commenges[1], Cécile Proust-Lima[1], Niall P. Murphy[2], Olivier Blin[3], Laurent Magy[4], Daniel Cohen[2], Shahram Attarian[5]

1 Inserm U1219, Université de Bordeaux, Bordeaux, Aquitaine, France, **2** Pharnext, Issy-Les-Moulineaux, France, **3** CIC-CPCET, Service de Pharmacologie Clinique et Pharmacovigilance, AP-HM, Aix Marseille Université, Marseille, France, **4** Centre de Référence Neuropathies Périphérique Rares, CHU de Limoges - Hôpital Dupuytren, Limoges, France, **5** Aix Marseille Université, INSERM, GMGF, Marseille, France

* wenjia.wang@pharnext.com

Citation: Wang W, Guedj M, Bertrand V, Foucquier J, Jouve E, Commenges D, et al. (2017) A Rasch Analysis of the Charcot-Marie-Tooth Neuropathy Score (CMTNS) in a Cohort of Charcot-Marie-Tooth Type 1A Patients. PLoS ONE 12(1): e0169878. doi:10.1371/journal.pone.0169878

Editor: Jong-Ling Fuh, Taipei Veterans General Hospital, TAIWAN

Received: September 29, 2016

Accepted: December 24, 2016

Published: January 17, 2017

Data Availability Statement: The restriction prohibiting the authors for making the minimal dataset publicly available is as follows: in the signed ICF, the participants were aware that the study result and not the dataset could be transmitted to scientific associations through congress and/or publication. However, the authors will analyse all demand from individual readers and will put in place efforts to best fulfil the request. Data are from studies CLN-PXT3003-01; BMK-CMT-01 and BMK-CMT-02 whose authors may be contacted at

Abstract

The Charcot-Marie-Tooth Neuropathy Score (CMTNS) was developed as a main efficacy endpoint for application in clinical trials of Charcot-Marie-Tooth disease type 1A (CMT1A). However, the sensitivity of the CMTNS for measuring disease severity and progression in CMT1A patients has been questioned. Here, we applied a Rasch analysis in a French cohort of patients to evaluate the psychometrical properties of the CMTNS. Overall, our analysis supports the validity of the CMTNS for application to CMT1A patients though with some limitations such as certain items of the CMTNS being more suitable for moderate to severe forms of the disease, and some items being disordered. We suggest that additional items and/or categories be considered to better assess mild-to-moderate patients.

Background

Charcot-Marie-Tooth (CMT) disease is the most common inherited disorder of the peripheral nervous system [1,2]. CMT type 1A (CMT1A), caused by a duplication of the myelin protein encoding gene *PMP22* [3,4], accounts for 50% of patients with CMT [1,2,5]. A typical feature of CMT1A is weakness of the foot and lower leg muscles, which may lead to foot drop and a high-stepped gait with frequent tripping or falls. Currently, there are no approved treatments for CMT1A disease though there have been considerable interest in the potential of ascorbic acid (AA) as a therapy leading to six clinical trials investigating the efficacy of AA on CMT1A [6–11]. Unfortunately, no beneficial clinical effects of AA were identified in any of these trials, as confirmed by two meta-analyses [12,13]. Recently, a clinical trial of PXT3003 (a fixed combination of baclofen, naltrexone and sorbitol) showed preliminary evidence of efficacy in an exploratory phase 2 study [14], which was also confirmed by a meta-analysis [12]. A conclusion shared by all of these studies is that selecting a clinically meaningful efficacy endpoint for CMT1A trials is challenging. Among the reasons for this are questions surrounding the relevance of efficacy endpoints, which remains an active topic of discussion. With regard to this,

mguedj@pharnext.com. The rest of data whose authors may be contacted at Olivier.Blin@aphm.fr.

Funding: The study was sponsored by Pharnext. The funder provided support in the form of salaries for authors WW, MG, VB, JF, NPM, DCoh, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing Interests: WW, MG, VB, JF, NPM and DCoh are employees of Pharnext. MG, VB, JF, NPM and DCoh have shares in the company. MG, VB and DCoh hold patent applications by Pharnext. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

the Charcot-Marie-Tooth Neuropathy Score (CMTNS) was first proposed and validated by Shy *et al.* to provide a reliable measure of impairment in CMT [15]. The CMTNS is composed of 9 items evaluating different functions related to the disease: 5 of impairment ('*Sensory Symptoms*', '*Pin Sensibility*', '*Vibration*', '*Strength Arms*' and '*Strength Legs*'), 2 of activity limitations ('*Motor Symptoms Arms*' and '*Motor Symptoms Legs*') and 2 electrophysiological measures ('*Ulnar CMAP*' and '*Ulnar SNAP*'). Each item is scored from 0 to 4 and the total sum of the item scores provides a global measure of disease severity, with higher scores indicating worsening function [15].

The CMTNS has been used as the primary or main endpoint in most completed clinical trials for CMT1A to date. However, the ability of the CMTNS to measure responses to treatment has not been demonstrated and among all the studies published, meta-analysis reveals significant improvements on the CMTNS only under PXT3003 versus placebo [12]. With this in mind, the sensitivity of the CMTNS to change and its psychometric properties are still debated. In particular, it has been suggested that some components of the CMTNS are too insensitive, mainly because of floor and ceiling effects [16]. Therefore, a modified version of the scale (CMTNS-v2) has been proposed by Murphy *et al.* [17] in an attempt to reduce the aforementioned effects and to standardize patient assessment. This version has also been questioned recently and a 'weighted' alternative has been suggested as a potential improvement [18]. Finally, a modified CMTNS (called CMTNS-Mod) has also been proposed by adding three functional measures (9-hole peg test, foot dorsiflexion and walk test) while removing four of the initial items ('*Ulnar SNAP*', '*Pin Sensibility*', '*Vibration*' and '*Strength of Arms*') [19]. However, none of these modified versions have been evaluated in natural history or therapeutic trials. As the CMTNS is the only CMT specific outcome measure available and has been widely applied, it is important to review its properties and find directions in which it could be improved. Firstly, it is important to demonstrate the sensitivity of CMTNS scores with disease progression. With regard to this, the CMTNS showed modest changes over time in longitudinal studies. Shy *et al.* reported a mean increase of 0.69 points per year for natural progression [20]. In parallel, clinical trials showed that CMT1A under placebo deteriorates even more slowly with a mean increase of 0.16 points per year [12].

To be a valid indicator of disease severity, the CMTNS should also comply with requirements of modern measurement theory such as unidimensionality (which implies that the scale measures only one construct, which allows the items to be summed together to form a scale with only one dimension), internal construct validity and reliability [21]. One well-accepted way to provide such evidence is to perform a Rasch model analysis [22], which has been widely employed in clinical scale construction and validation [21,23,24]. The Rasch Model assesses a latent trait, such as disease severity, by the responses of patients to a set of items [25]. It provides a range of diagnostic information that can be used to determine how well each item contributes to the measurement of the latent trait and doing so, it helps in assessing the validity of the scale and its possible axes for improvement.

Here, we performed a Rasch analysis of the CMTNS in a cohort of 277 mild-to-severe CMT1A patients from merging of two French clinical trials [9,14] and one non-investigational study.

Methods

Participants and setting

CMT1A patients involved in this study initially participated in the French phase 2 clinical trial of ascorbic acid led by Micallef *et al.* [9] and/or in the phase 2 clinical trial of PXT3003 led by Attarian *et al.* [14] and/or in a subsequent non-investigational clinical study (BMK-CMT)

sponsored by Pharnext. The ethics committee "Comité de Protection des Personnes Sud-Méditerranée I" has approved this study and the RCB ID is 2010-023097-40. Participants provide their written informed consent to participate in the study. Patients were included from six hospital sites in France: Marseille, Lille, Limoges, Lyon, Nantes and Paris. A total of 277 patients completed the CMTNS scoring in at least one of the 3 aforementioned clinical studies. Among them, there were 110 men and 167 women, with ages ranging from 18 to 69 and an average age of 45. Patient CMTNS scores ranged from 2 to 31 and were classified into mild (CMTNS $\leq$ 10, 47 patients), moderate (11 $\leq$ CMTNS $\leq$ 20, 201 patients) or severe (CMTNS $\geq$ 21, 29 patients).

The Rasch model

The Rasch model is a mathematical framework initially proposed to analyze rating scales and evaluates a latent variable not measurable directly from a set of categorical items (eg, disability, cognition or quality of life). In this model, the raw score of each item is transferred into interval scaling by a logistic function where data are found to meet the model assumptions. Both the person's ability and the item difficulty, also referred to as person and item parameters, are defined on the same dimension. If a person's ability is known, it is possible to predict how that person is likely to perform on a given item. Specifically, the probability of a response is modeled as a logistic function of the difference between the person and the item parameters. The Rasch model was initially developed for dichotomous items, and then extended to polytomous items in which successive integer scores represent categories of increasing level of disability such as the CMTNS.

Rasch model analysis

The Rasch model analysis provides an integrated framework for evaluating if a sum score (such as the CMTNS) satisfies a set of requirements listed hereafter. They first include the three assumptions of the Rasch Model, that of local independence, unidimensionality and invariance, briefly explained as follows:

Local independence means that, conditionally on the latent person ability, the response of a particular individual to an item depends neither on the responses to other items nor on the responses given by other people to the same item. This is examined by the residual correlations between items, which should be no more than 0.3 for each pair of items [26].

Unidimensionality. The Rasch model assumes that the response to each item depends on a unique latent trait. It can be assessed by creating two subsets of items using a Principal Component Analysis (PCA) of the item residuals, with those loading negatively forming one set, and those loading positively forming the second set. Each person parameter estimated from one set of items is then compared to those derived from the other set of items using a *t*-test. If less than 5% of these tests are significant at the 5% level, then unidimensionality is supported [27]. Another approach to examine unidimensionality is to apply a generalization of the Martin-Löf test to the two subsets of items defined previously [28]. A non-significant *p*-value for this test at the 5% level supports the assumption of unidimensionality.

Invariance means that item difficulties remain the same across different groups, such as age or gender. The invariance of items is assessed through an analysis of variance of the residuals where the key group of interest is the main factor. If the inter-person-group variance is statistically significant, the item bias is called Differential Item Functioning [29,30]. When it is present, the probability of an item response cannot be explained totally by the person and item parameters, as it is also influenced by other group properties such as age and gender. Here,

each item was checked for Differential Item Functioning across two subgroups: gender (male and female), age (younger or older than 45 years).

Once the three assumptions of local independence, unidimensionality and invariance are met, it is possible to use the Rasch model to further evaluate the scale by investigating overall goodness-of-fit, reliability, fitness of individuals or items, and consistency of items, as introduced following:

Overall goodness-of-fit. Most publications dealing with Rasch analysis estimate the overall goodness-of-fit using a chi-square test [22,23,31]. If the data fit the Rasch Model, a summary chi-square interaction statistic should be non-significant. However, recent studies show that chi-square approaches are problematic: these indices are too powerful and the appropriate degree of freedom is often not clear [32,33]. Instead, the Andersen's likelihood-ratio test [34] shows high power and acceptable type-I error rate in Rasch Model estimation [35]. To perform this test, subjects are split into $g = 1, \ldots, G$ score-level subgroups in which a conditional likelihood is computed and compared to the total conditional likelihood computed in the complete sample of subjects. The statistic of the test is given by:

$$2(\Sigma_{g=1}^{G} logL_C^{(g)} - logL_C),$$

where $L_C^{(g)}$ is the conditional likelihood of subgroup g and L_C is the total conditional likelihood. This statistic has an asymptotic chi-square distribution with degrees of freedom equal to the number of parameters estimated in the score groups minus the number of parameters estimated in the complete data set. A non-significant p-value for this test indicates goodness-of-fit for the Rasch model.

Reliability of the CMTNS scale is estimated by the Person Separation Index (PSI) given by the proportion of true variance relative to the true and error variance. In practice it measures the internal consistency and the discrimination power of the scale, i.e. the ability of the scale to discriminate amongst persons with different levels of the trait. It is equivalent to the Cronbach's alpha [36], but it uses the person estimates in logits instead of the raw scores. A PSI value greater than 0.7 is considered as acceptable.

Item fit can be assessed by several indicators. The residual item fit statistics are expected to approximate a Normal distribution (mean close to 0 with a SD close to 1), which is tested using a chi-square test [21]. A significant chi-square test based p-value may indicate misfit. In parallel, a similar analysis could be performed for the test of person fit. Then, fit statistics can be computed and focus on two aspects: infit (means inlier-sensitive fit) and outfit (means outlier-sensitive fit). Infit is more sensitive to the overall pattern and less influenced by outliers and thus infit problems are more of a threat to measurement than outfit ones. Infits and outfits are reported in both mean squares and standardized fit t-statistics. The mean squares indicate the amount of distortion of the measurement system whereas the t-statistics indicate how likely the item is misfit [37]. Mean-squares greater than 1.3 indicates underfit to the Rasch model, i.e., the data are less predictable than the model expects; mean-squares less than 0.7 indicate overfit to the Rasch model, i.e., the data are more predictable than the model expects. High t-statistics (> 2.0) show that the item distorts or degrades the measurement system as underfit while low t-statistics (< -2.0) mean data are too predictable or overfit, but not degrading. Underfit and overfit to the model have different implications for measurement. Underfit degrades the quality of the measurement and should prompt reflection on its cause. Overfit might mislead one into concluding that the quality of the measure is better than it really is, and has less practical implication than underfit [38].

Consistency of items. A particularly useful output of the Rasch analysis is the person-item map (also sometimes referred to as 'Wright map'). This map displays the difficulty of the

items on the same latent dimension as the impairment of the patients. For each item, a threshold of a category is defined as the location at which the cumulative probability of selecting this category versus all the other options reaches 0.5. In doing so, thresholds should follow the same order as categories. A disorder of categories in an item occurs when the ordinal numbering of categories is not in accord with their fundamental meaning or when individuals have difficulties in consistently discriminating categories. In this case, the disordered categories should be rearranged and Item Characteristic Curves representing the probability of selecting each category for one item can be plotted in order to examine whether this disorder item from under or over-selection of one category.

Implementation

The Rasch Model has various mathematical variations. Here, we precisely considered the Partial Credit Model [37] allowing different response format for each item, which is the case of the CMTNS. A more detailed introduction to the Partial Credit Model can be found in Wang *et al* [39]. Analyses were performed with *R* (http://cran.r-project.org). The dimensionality, local dependency and invariance analyses were carried out using custom-made *R* functions, while the other Rasch analyses were performed with the *R* package *eRm* [40]. Statistical significance was considered at the 5% level and Bonferroni correction for multiple testing was applied where appropriate.

Results

We performed a Rasch analysis of the CMTNS using responses from the 277 individuals included in the study. A well targeted sample size of at least 150 individuals is required to reach a 99% confidence that the estimated item difficulty is within +/-0.5 logit of its stable value [41]. Our sample of 277 CMT1A patients was therefore adequate for the analysis. A preliminary quality control based on a significant *p*-value of the person fit chi-square test excluded 15 individuals (5.4%). From there, 262 individuals were included in the Rasch analysis.

Local independency, unidimensionality and invariance

We investigated the compliance of the CMTNS to the main assumptions of the Rasch model. Firstly, local independency was shown by the absence of pairwise correlations between item fit residuals greater than 0.3 (Fig 1). Then, unidimensionality was supported by the fact that only 2 patients of 262 total (much less than 5%) had a significant *p*-value following the PCA approach described in the methods. The *p*-value of the Martin-Löf test was not significant ($p = 0.919$). Finally, the response residuals of different subgroups (gender, age) in each item do not display significant Differential Item Functioning, which means invariance of items. These results led us to conclude that the CMTNS meet the assumptions of the Rasch Model in our cohort of CMT1A patients.

Overall goodness-of-fit and reliability

A non-significant *p*-value of the Andersen's likelihood-ratio test ($p = 0.435$) indicates a good overall fit of the CMTNS to the Rasch model. The PSI calculated on our data equals 0.715, pointing to acceptable reliability of the CMTNS, although this value is not particularly high.

Item fit

On the item level, '*Ulnar SNAP*' is the only item of the CMTNS that has a significant chi-square based *p*-value at the 5% level ($p = 0.044$). However, it is not significant after

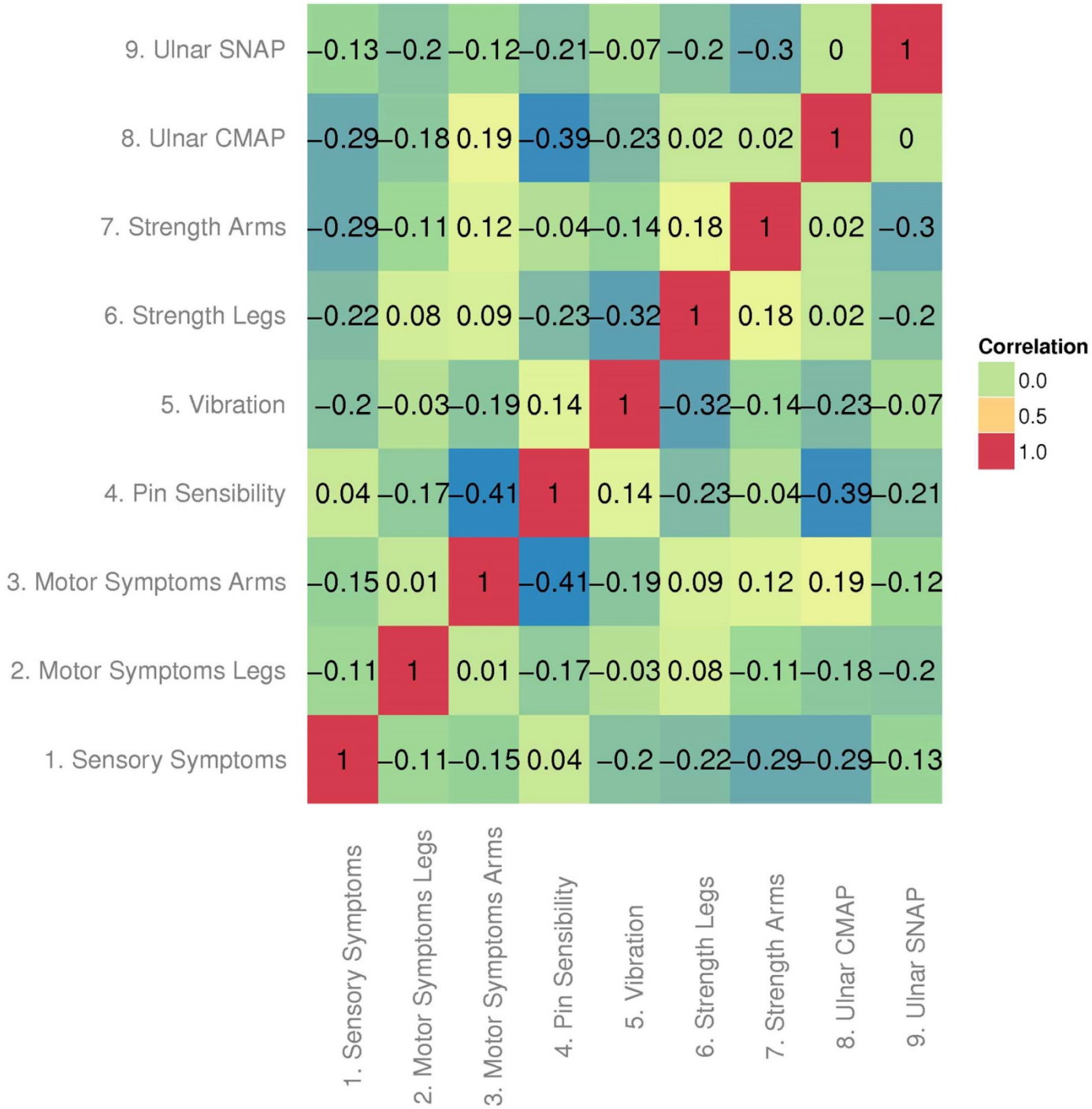

Fig 1. Pairwise correlations between item fit residuals. Pearson correlation coefficients between the residuals of 9 items in the CMTNS. CMAP = Amplitudes of Compound Muscle Action Potentials; SNAP = Amplitudes of Sensory Nerve Action Potentials.

doi:10.1371/journal.pone.0169878.g001

Bonferroni correction. Residuals of all items have a distribution with means close to 0 and SD close to 1 (see Table 1). None of the infit *t* statistics are superior to 2 (Fig 2), which is to say no item distorts the measurement. Both infit and outfit of the '*Strength Legs*' and '*Strength Arms*' items are inferior to -2, and the fit mean squares of '*Strength Legs*' was slightly lower

Table 1. Item statistics. MSQ = Mean-square.

Item names	Location	Residual mean	Residual SD	*p*-value	Outfit MSQ	Infit MSQ	Outfit *t*-stat	Infit *t*-stat
1. Sensory Symptoms	0.806	0.001	1.049	0.128	1.096	1.138	0.964	1.491
2. Motor Symptoms Legs	0.942	-0.031	0.924	0.958	0.851	0.862	-1.675	-1.699
3. Motor Symptoms Arms	1.54	-0.016	0.931	0.941	0.864	0.861	-1.616	-1.597
4. Pin Sensibility	0.429	0.001	0.943	0.902	0.885	0.856	-1.431	-1.88
5. Vibration	0.353	-0.014	0.966	0.775	0.929	0.941	-0.793	-0.69
6. Strength Legs	0.105	-0.034	0.81	1	0.654	0.661	-4.689	-4.789
7. Strength Arms	0.755	-0.008	0.864	0.999	0.744	0.745	-3.28	-3.251
8. Ulnar CMAP	0.687	0.010	1.031	0.23	1.059	1.007	0.76	0.116
9. Ulnar SNAP	-0.788	0.018	1.074	0.044	1.15	1.054	0.86	0.55

doi:10.1371/journal.pone.0169878.t001

than 0.7, which points to responses to the two items as being too predictable, possibly leading to overfit.

Consistency of items

The person-item map (Fig 3) displays the location of person abilities and item difficulties respectively along the same latent dimension. Although the category thresholds of most items cover mild-to-severe range of disability well, item difficulty locations clump at the range of patients with higher person parameters (right side of the latent dimension), which means that they have more probability to differentiate patients with higher level of disease severity. For instance, '*Motor Symptom Arm*' shows the highest item difficulty meaning that mild-to-moderate patients are more likely to answer '0' (i.e. no disability in arms) for this item. Three items have disordered categories ('*Sensory Symptoms*', '*Motor Symptoms Legs*' and '*Ulnar SNAP*') indicated in red on the person-item map (Fig 3). To further investigate these disordered items, we examined the Item Characteristic Curves (Fig 4). Category 2 in '*Motor Symptoms Legs*' (*i.e.* ankle-foot orthosis on at least one leg or ankle support) was under-selected, which causes the observed disorder. '*Sensory Symptoms*' and '*Ulnar SNAP*' have the categories 0 and 4 evidently over-selected compared to other categories, which means that they are not adapted to discriminate CMT1A patients well.

Proposed modification of the CMTNS

In attempts to improve item fit to the model, a common strategy is to collapse adjacent categories when they have disordered thresholds. Given our results, we collapsed Categories 2, 3 and 4 into one category in '*Sensory Symptoms*' and '*Ulnar SNAP*' and Categories 2, 3 in '*Motor Symptoms Legs*'. The person-item map of the modified data shows that all items are now well-ordered (Fig 5). However, after these modifications, the PSI of the CMTNS does not improve (= 0.713 now), and the infit *t*-statistics of '*Sensory Symptoms*' increased from 1.49 to 1.88. Although item categories are well ordered after our modifications, this modification does not enhance the overall fitness of the CMTNS to the Rasch Model.

Discussion

The CMTNS was developed by Shy *et al.* [15] as the first composite clinical scale dedicated to quantifying impairment and measuring progression in CMT patients. Although the validity of CMTNS to assess severity has never been questioned, its sensitivity to change and its ability measure a response to treatment are still debated. Subsequent versions of the CMTNS have been proposed, such as the CMTNS-v2 by Murphy *et al.* [42] to attempt to reduce floor and

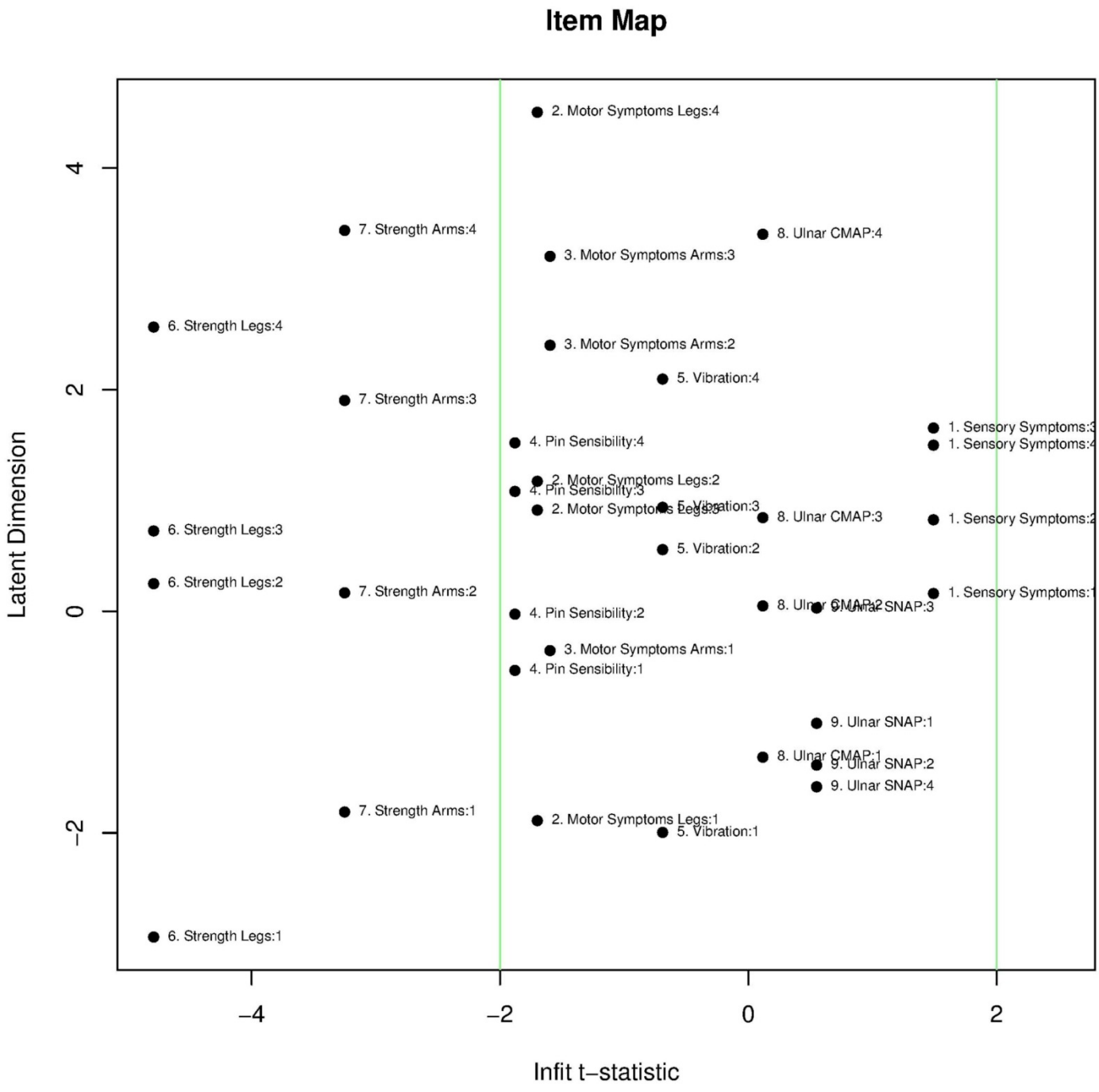

Fig 2. Infit statistics of categories of each item. The *x* axis displays the infit *t*-statistics of each category in 9 items. The *y* axis represents category distributions on the logit scale of latent dimension. Most items fall within the range -2 to 2 (indicated by green lines). CMAP = Amplitudes of Compound Muscle Action Potentials; SNAP = Amplitudes of Sensory Nerve Action Potentials.

doi:10.1371/journal.pone.0169878.g002

ceiling effects, the CMTNS-Mod by Mannil *et al.* [19] by adding three functional measures while removing three of the initial items, and finally a 'weighted' alternative of the CMTNS-v2 by Sadjadi *et al.* [18] resulting from a Rasch analysis. None of these modified versions have been evaluated in natural history or therapeutic trials.

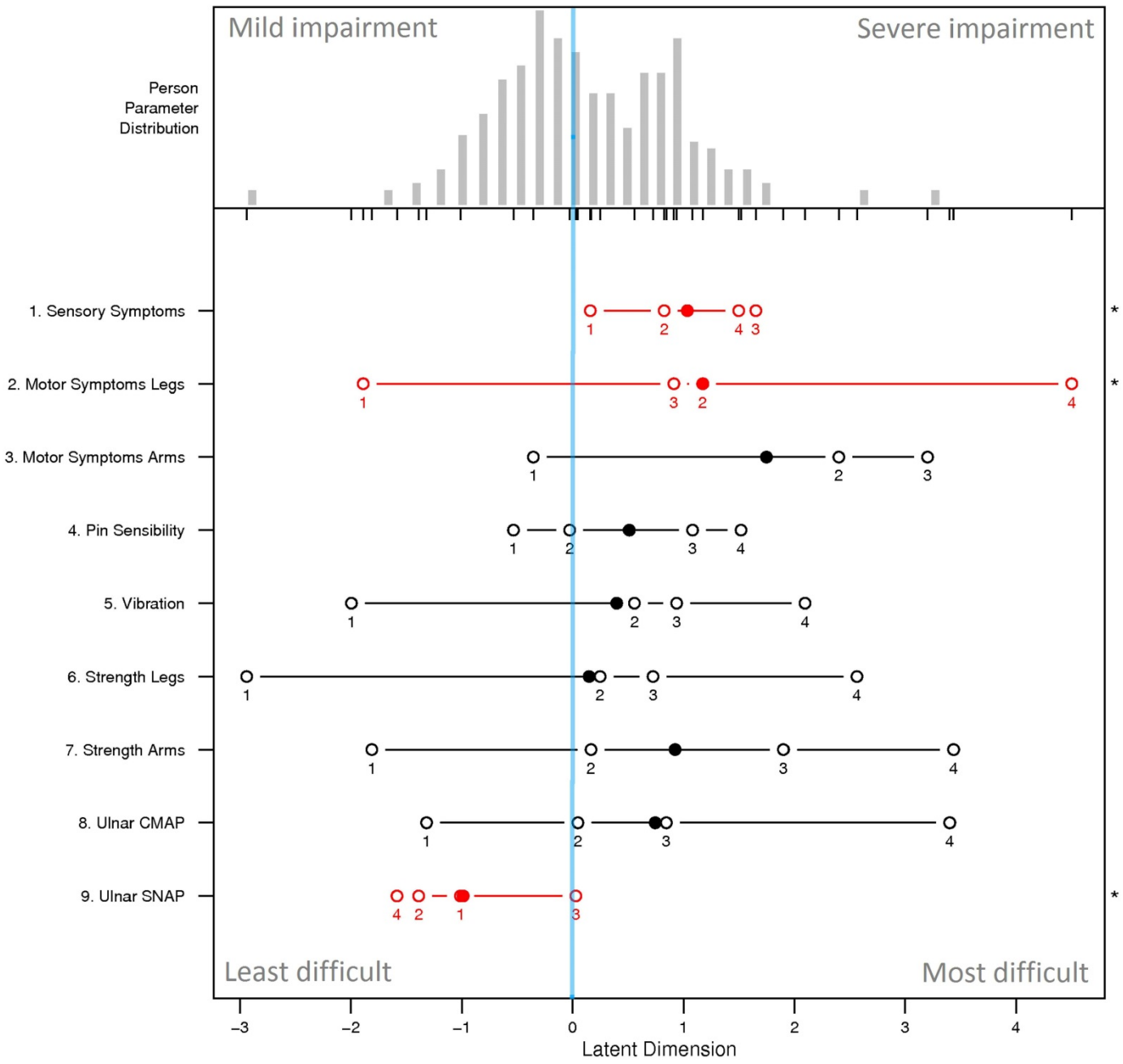

Fig 3. Person-Item map. The person-item map displays the location of person abilities and item difficulties respectively along the same latent dimension. The person parameter is located on the scale from left (mild impairment) to right (severe impairment). Locations of item difficulties are displayed with solid circles and thresholds of adjacent category locations with open circles. Items with disordered thresholds are marked in red and asterisks. The blue line indicates the zero level of the latent trait.

doi:10.1371/journal.pone.0169878.g003

In order to further investigate some key properties of the original CMTNS and to identify possible directions of improvement, we performed a validation of this scale based on a Rasch analysis in a cohort of 277 mild-to-severe CMT1A French patients, made possible by the integration of 3 studies including 2 published clinical trials [9,14]. Our first result is that overall and in the context of the Rasch analysis, the CMTNS appears as a valid measurement for

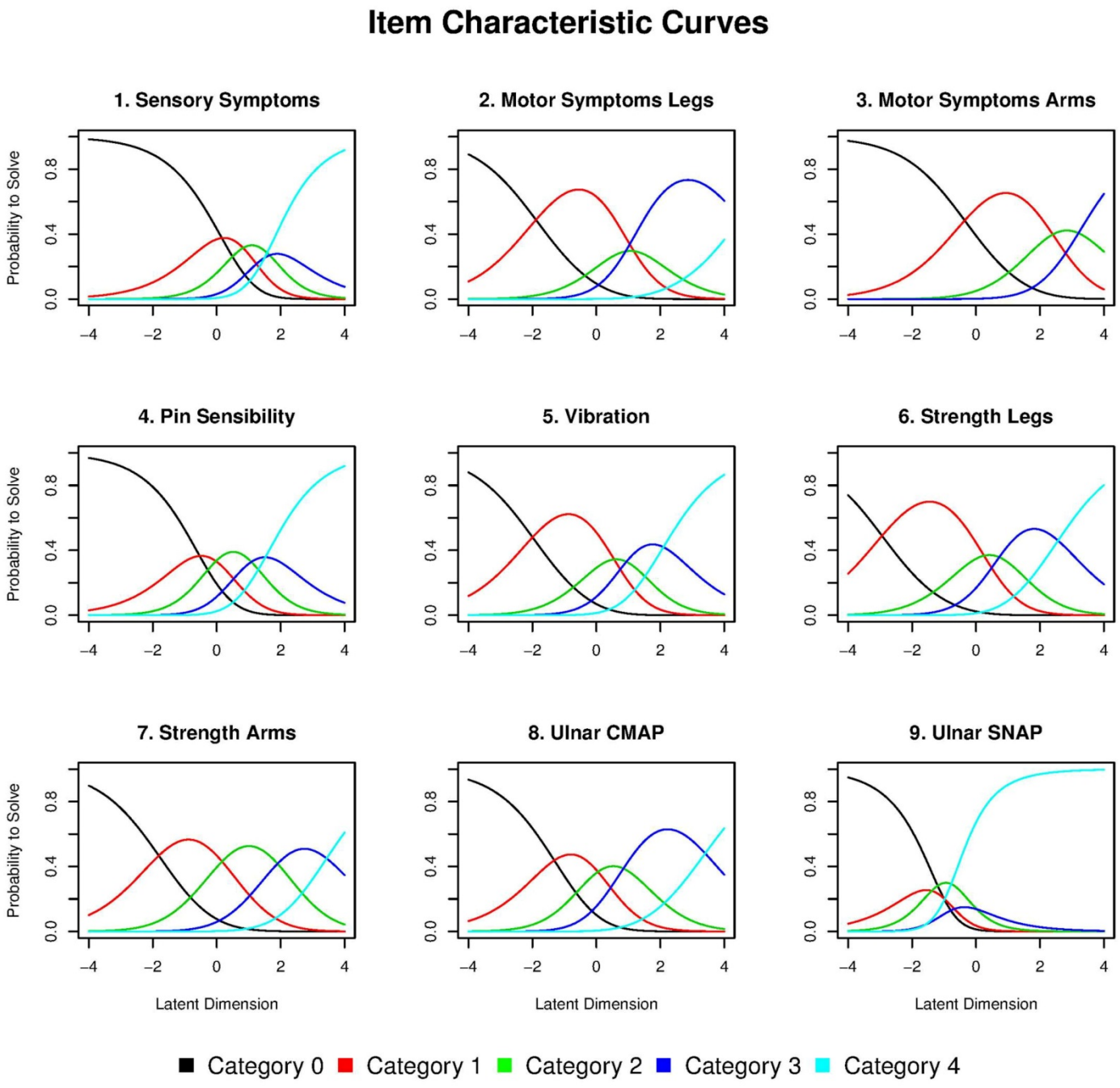

Fig 4. Item Characteristic Curves. Curves represent the probability of selecting a category in an item.

doi:10.1371/journal.pone.0169878.g004

CMT1A: the three main assumptions of the Rasch model (local independency, unidimensionality and invariance) were met, the scale showed good overall fit to the Rasch model and an acceptable measure of reliability. When analyzed individually, only two items ('*Strength Legs*' and '*Strength Arms*') showed an overfit to the model (infit and outfit t-statistics < -2), which has little major implication for the quality of the measurement. In the Rasch analysis of Sadjadi *et al.* [18] of the CMTNS-v2, all of the items showed good fit supporting the idea that they belong in the scale and contribute to the overall score of impairment. Although the CMTNS-

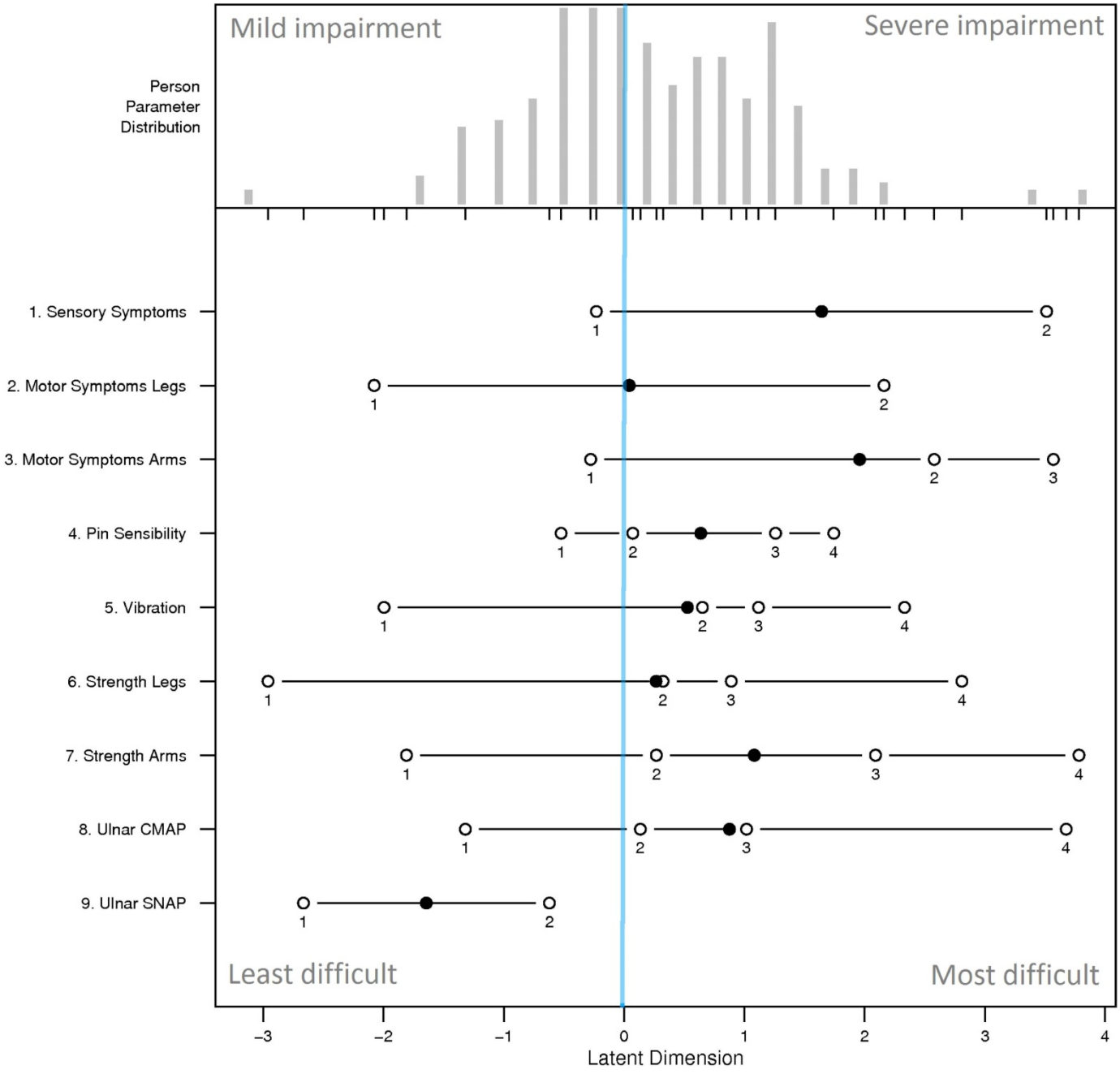

Fig 5. Person-Item map of the CMTNS after modification.

doi:10.1371/journal.pone.0169878.g005

v2 presents some modifications to the CMTNS in terms of categories or instruments of measure, they are very similar and can be discussed together here.

As a limitation, the person-item map suggests that the items of the CMTNS are more suitable for assessing moderate to severe forms of the disease, with the exception of '*Ulnar SNAP*'. Sadjadi *et al.* [18] arrived at the same conclusion, though in the CMTNS-v2, SNAP is measured on the radial nerve instead of the ulnar nerve, underlining the consistency of this result. This finding is also supported by a comparison study of the CMTNS-v2 and a pediatric version of

6. Verhamme C, de Haan R, Vermeulen M, Baas F, de Visser M, van Schaik I. Oral high dose ascorbic acid treatment for one year in young CMT1A patients: a randomised, double-blind, placebo-controlled phase II trial. {BMC} Med. 2009; 7: 70. doi: 10.1186/1741-7015-7-70 PMID: 19909499
7. Burns J, Ouvrier RA, Yiu EM, Joseph PD, Kornberg AJ, Fahey MC, et al. Ascorbic acid for Charcot-Marie-Tooth disease type 1A in children: a randomised, double-blind, placebo-controlled, safety and efficacy trial. Lancet Neurol. 2009; 8: 537–544. doi: 10.1016/S1474-4422(09)70108-5 PMID: 19427269
8. Toth C. Poor tolerability of high dose ascorbic acid in a population of genetically confirmed adult Charcot-Marie-Tooth 1A patients. Acta Neurologica Scandinavica. 2009. pp. 134–138. doi: 10.1111/j.1600-0404.2008.01134.x PMID: 19154534
9. Micallef J, Attarian S, Dubourg O, Gonnaud P-M, Hogrel J-Y, Stojkovic T, et al. Effect of ascorbic acid in patients with Charcot-Marie-Tooth disease type 1A: a multicentre, randomised, double-blind, placebo-controlled trial. Lancet Neurol. 2009; 8: 1103–1110. doi: 10.1016/S1474-4422(09)70260-1 PMID: 19818690
10. Pareyson D, Reilly MM, Schenone A, Fabrizi GM, Cavallaro T, Santoro L, et al. Ascorbic acid in Charcot-Marie-Tooth disease type 1A (CMT-TRIAAL and CMT-TRAUK): a double-blind randomised trial. Lancet Neurol. 2011; 10: 320–8. doi: 10.1016/S1474-4422(11)70025-4 PMID: 21393063
11. Lewis RA, McDermott MP, Herrmann DN, Hoke A, Clawson LL, Siskind C, et al. High-dosage ascorbic acid treatment in Charcot-Marie-Tooth disease type 1A: results of a randomized, double-masked, controlled trial. JAMA Neurol. 2013; 70: 981–7. doi: 10.1001/jamaneurol.2013.3178 PMID: 23797954
12. Mandel J, Bertrand V, Lehert P, Attarian S, Magy L, Micallef J, et al. A meta-analysis of randomized double-blind clinical trials in CMT1A to assess the change from baseline in CMTNS and ONLS scales after one year of treatment. Orphanet J Rare Dis. 2015; 10: 74. doi: 10.1186/s13023-015-0293-y PMID: 26070802
13. Gess B, Baets J, De Jonghe P, Reilly MM, Pareyson D, Young P. Ascorbic acid for the treatment of Charcot-Marie-Tooth disease. Cochrane database Syst Rev. 2015; 12: CD011952.
14. Attarian S, Vallat J-M, Magy L, Funalot B, Gonnaud P-M, Lacour A, et al. An exploratory randomised double-blind and placebo-controlled phase 2 study of a combination of baclofen, naltrexone and sorbitol (PXT3003) in patients with Charcot-Marie-Tooth disease type 1A. Orphanet J Rare Dis. 2014; 9: 199. doi: 10.1186/s13023-014-0199-0 PMID: 25519680
15. Shy ME, Blake J, Krajewski K, Fuerst DR, Laura M, Hahn AF, et al. Reliability and validity of the CMT neuropathy score as a measure of disability. Neurology. 2005; 64: 1209–14. doi: 10.1212/01.WNL.0000156517.00615.A3 PMID: 15824348
16. Reilly MM, Shy ME, Muntoni F, Pareyson D. 168th ENMC International Workshop: outcome measures and clinical trials in Charcot-Marie-Tooth disease (CMT). Neuromuscul Disord. 2010; 20: 839–46. doi: 10.1016/j.nmd.2010.08.001 PMID: 20850975
17. Murphy SM, Herrmann DN, McDermott MP, Scherer SS, Shy ME, Reilly MM, et al. Reliability of the CMT neuropathy score (second version) in Charcot-Marie-Tooth disease. J Peripher Nerv Syst. 2011; 16: 191–198. doi: 10.1111/j.1529-8027.2011.00350.x PMID: 22003934
18. Sadjadi R, Reilly MM, Shy ME, Pareyson D, Laura M, Murphy S, et al. Psychometrics evaluation of Charcot-Marie-Tooth Neuropathy Score (CMTNSv2) second version, using Rasch analysis. J Peripher Nerv Syst. 2014; 19: 192–6. doi: 10.1111/jns.12084 PMID: 25400013
19. Mannil M, Solari A, Leha A, Pelayo-Negro AL, Berciano J, Schlotter-Weigel B, et al. Selected items from the Charcot-Marie-Tooth (CMT) Neuropathy Score and secondary clinical outcome measures serve as sensitive clinical markers of disease severity in CMT1A patients. Neuromuscul Disord. 2014; 24: 1003–17. doi: 10.1016/j.nmd.2014.06.431 PMID: 25085517
20. Shy ME, Chen L, Swan ER, Taube R, Krajewski KM, Herrmann D, et al. Neuropathy progression in Charcot-Marie-Tooth disease type 1A. Neurology. 2008; 70: 378–383. doi: 10.1212/01.wnl.0000297553.36441.ce PMID: 18227419
21. Kersten P, White PJ, Tennant A. Is the pain visual analogue scale linear and responsive to change? An exploration using rasch analysis. PLoS One. 2014; 9.
22. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care Res. 2007; 57: 1358–1362.
23. Zucca AC, Lambert SD, Boyes AW, Pallant J. Rasch analysis of the Mini-Mental Adjustment to Cancer Scale (mini-MAC) among a heterogeneous sample of long-term cancer survivors: A crosssectional study. Health and Quality of Life Outcomes. 2012. p. 55. doi: 10.1186/1477-7525-10-55 PMID: 22607052
24. van Nes SI, Vanhoutte EK, van Doorn P a, Hermans M, Bakkers M, Kuitwaard K, et al. Rasch-built Overall Disability Scale (R-ODS) for immune-mediated peripheral neuropathies. Neurology. 2011; 76: 337–45. doi: 10.1212/WNL.0b013e318208824b PMID: 21263135

25. Fischer GH. Rasch Models [Internet]. Fischer GH, Molenaar IW, editors. New York, NY: Springer New York; 1995.
26. Andrich D. Sufficiency and conditional estimation of person parameters in the Polytomous Rasch Model. Psychometrika. 2010; 75: 292–308.
27. Smith E V. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas. 2002; 3: 205–231. PMID: 12011501
28. Christensen KB, Bjorner JB, Kreiner S, Petersen JH. Testing unidimensionality in polytomous Rasch models. Psychometrika. 2002. pp. 563–574.
29. Hanson BA. Uniform DIF and DIF Defined by Differences in Item Response Functions. Journal of Educational and Behavioral Statistics. 1998. pp. 244–253.
30. Kim SH, Cohen AS, Alagoz C, Kim S. DIF detection and effect size measures for polytomously scored items. J Educ Meas. 2007; 44: 93–116.
31. Burns J, Ouvrier R, Estilow T, Shy R, Laurá M, Pallant JF, et al. Validation of the Charcot-Marie-Tooth disease pediatric scale as an outcome measure of disability. Ann Neurol. 2012; 71: 642–52. doi: 10.1002/ana.23572 PMID: 22522479
32. Orlando M, Thissen D. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. Applied Psychological Measurement. 2000. pp. 50–64.
33. Orlando M, Thissen D. Further Investigation of the Performance of S—X2: An Item Fit Index for Use With Dichotomous Item Response Theory Models. Applied Psychological Measurement. 2003. pp. 289–298.
34. Andersen EB. A goodness of fit test for the rasch model. Psychometrika. 1973; 38: 123–140.
35. Suárez-Falcón JC, Glas CAW. Evaluation of global testing procedures for item fit to the Rasch model. Br J Math Stat Psychol. 2003; 56: 127–143. doi: 10.1348/000711003321645395 PMID: 12803827
36. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16: 297–334.
37. Masters GN. A Rasch model for partial credit scoring. Psychometrika. Springer; 1982; 47: 149–174. http://www.springerlink.com/index/5000VV4281622815.pdf
38. Green K, Frantom C. Survey development and validation with the Rasch model. Int Conf Quest Dev Eval Test. 2002; 1–30.
39. Wang W, Mandel J, Bouaziz J, Commenges D, Nabirotchkine S, Chumakov I, et al. A Multi-Marker Genetic Association Test Based on the Rasch Model Applied to Alzheimer's Disease. PLoS One. Public Library of Science; 2015; 10: e0138223.
40. Mair P, Hatzinger R. Extended Rasch Modeling: The R Package eRm. 2007;
41. Linacre JM. WINSTEPS Rasch measurement computer program. Chicago Winsteps com. 2006;
42. Murphy SM, Herrmann DN, McDermott MP, Scherer SS, Shy ME, Reilly MM, et al. Reliability of the CMT neuropathy score (second version) in Charcot-Marie-Tooth disease. J Peripher Nerv Syst. 2011; 16: 191–8. doi: 10.1111/j.1529-8027.2011.00350.x PMID: 22003934
43. Burns J. Transitioning outcome measures: relationship between the CMTPedS and CMTNSv2 in children, adolescents, and young adults with Charcot-Marie-Tooth disease. J Peripher Nerv Syst. 2013; 18, 2.

Chapter 3

Rasch Based Genetic Association Test

3.1. Introduction

In the previous chapter, Rasch model is applied to the CMTNS to evaluate the psychometric properties of this scale. Since GWAS data, which are widely used for neurodegenerative diseases, are also categorical data, we had the idea that this alternative psychometric method of the classical test theory may be helpful for GWAS data analysis. Alzheimer's disease is a complex disease and has no efficacious treatment. Determining the genes associated with the disease can facilitate the understanding of the pathological mechanism and furthermore treatment development. In a general way, there are multiple SNPs in a gene. The main question hence is how to summarize the association between multiple SNPs and a trait of interest into a single statistic. In this study, we utilized the Rasch model as a multi-marker genetic association test to find Alzheimer's disease associated genes.

This result is published on *Plos One* with the title "A Multi-Marker Genetic Association Test

Based on the Rasch Model Applied to Alzheimer's Disease".

3.2. Methods and Results

A set of SNPs which can be genotyped into 0, 1 and 2 can be considered as a rating scale with polytomous items. Therefore, we can apply Rasch model on each gene comprising of multiple SNPs. Rasch model provides an estimation of a person's location on the latent trait continuum (person parameter). By comparing the person parameters of the case group and control group, we could assess the association between a given gene and the disease of interst. The polytomous

Rasch model was implemented in *R* package *eRm*. The association was assessed by logistic model.

To evaluate this multi-marker genetic association test based on the Rasch model, we simulated a series of SNPs data and compared this method with four existing association tests. By comparing false-positive rate and power, we found the proposed approach showed good performances: it has correct false positive rate and high power.

This Rasch model based association test was applied to the GWAS data in the ADNI study to explore disease-associated genes. Among the top genes selected by the proposed method, several can be functionally linked to Alzheimer's disease. A pathway analysis of these genes also highlights the metabolism of cholesterol that is known to play a key role in AD pathogenesis. Moreover, these elements can be integrated into a hypothetic signalling network potentially targeted by a drug that shows efficacy in disease models.

3.3. Manuscript

PLOS | ONE

RESEARCH ARTICLE

A Multi-Marker Genetic Association Test Based on the Rasch Model Applied to Alzheimer's Disease

Wenjia Wang[1,2]*, Jonas Mandel[1], Jan Bouaziz[1], Daniel Commenges[2], Serguei Nabirotchkine[1], Ilya Chumakov[1], Daniel Cohen[1], Mickaël Guedj[1], the Alzheimer's Disease Neuroimaging Initiative[¶]

1 Pharnext, Issy-les-Moulineaux, Ile de France, France, **2** Inserm U897, University of Bordeaux, Bordeaux, Aquitaine, France

¶Membership of the Alzheimer's Disease Neuroimaging Initiative can be found in the Acknowledgments section.

* wenjia.wang@pharnext.com

OPEN ACCESS

Citation: Wang W, Mandel J, Bouaziz J, Commenges D, Nabirotchkine S, Chumakov I, et al. (2015) A Multi-Marker Genetic Association Test Based on the Rasch Model Applied to Alzheimer's Disease. PLoS ONE 10(9): e0138223. doi:10.1371/journal.pone.0138223

Editor: Momiao Xiong, University of Texas School of Public Health, UNITED STATES

Received: March 30, 2015

Accepted: August 27, 2015

Published: September 17, 2015

Data Availability Statement: All GWAS files were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and a public and freely accessible.

Funding: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). Pharnext provided support in the form of salaries for authors WW, MG, JM, JB, D. Cohen, SN, and IC but did not have any additional role in the

Abstract

Results from Genome-Wide Association Studies (GWAS) have shown that the genetic basis of complex traits often include many genetic variants with small to moderate effects whose identification remains a challenging problem. In this context multi-marker analysis at the gene and pathway level can complement traditional point-wise approaches that treat the genetic markers individually. In this paper we propose a novel statistical approach for multi-marker analysis based on the Rasch model. The method summarizes the categorical genotypes of SNPs by a generalized logistic function into a genetic score that can be used for association analysis. Through different sets of simulations, the false-positive rate and power of the proposed approach are compared to a set of existing methods, and shows good performances. The application of the Rasch model on Alzheimer's Disease (AD) ADNI GWAS dataset also allows a coherent interpretation of the results. Our analysis supports the idea that *APOE* is a major susceptibility gene for AD. In the top genes selected by proposed method, several could be functionally linked to AD. In particular, a pathway analysis of these genes also highlights the metabolism of cholesterol, that is known to play a key role in AD pathogenesis. Interestingly, many of these top genes can be integrated in a hypothetic signalling network.

Introduction

With the recent improvement of high-throughput genotyping technologies, the use of Genome-Wide Association Studies (GWAS) has become widespread in genetic research to identify significant associations between genetic markers such as Single Nucleotide Polymorphisms (SNPs) and complex phenotypes such as common diseases. GWAS generally yield results at the SNP-level, that are sets of SNPs associated with the disease. However, the vast

study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the "author contributions" section.

Competing Interests: The authors of this manuscript have read the journal's policy and have the following competing interests: WW, MG, JM, JB, D. Cohen, SN, and IC are paid employees in Pharnext. This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials.

majority of loci that have been identified for common diseases show modest effects and generally explain only a small part of the variance or heritability of the phenotype observed [1]. In a recent study of Body Mass Index (BMI), the markers associated explained only 0.84% of the variance, although it is considered that genetic factors should actually account for 40%-70% of the variance of BMI [2]. One explanation for the missing heritability is that the common analysis approach, assessing the effect of each SNP individually, is not well suited for the detection of small effects of multiple SNPs. Disease susceptibility is actually likely to depend on the cumulative effect of multiple variants in several genes interacting in functional pathways [3].

It is increasingly recognized that analyzing the combined association of multiple markers at the gene or pathway level may provide a complementary approach to the more common single SNP association approach, with several key benefits [4]. First it incorporates a priori biological knowledge in the analysis: as a matter of fact, in Genetics, the gene is often considered as the unit of interest since the analyses of the functional mechanisms of a disease are generally based on genes and their products such as RNA or proteins [5]. Determining the genes associated with the disease opens the door to a lot of additional research such as targeting genes of interests for candidate-gene studies or replicate association studies. Also, it allows the consideration of biological information, such as pathways or protein interactions, in the analysis of GWAS [6]. For instance, enrichment analysis such as performed by the method Gene Set Enrichment Analysis (GSEA) [7] aims to determine sets of genes involved in common biological processes or biological pathways. Such an analysis is possible through the use of functional information that is only available at the gene level. Second, as the number of genes or pathways is substantially smaller than the number of markers genotyped in GWAS, fewer hypotheses will be tested requiring less stringent multiple-testing correction [8]. Finally, by combining SNPs with modest associations, evidence of association at the gene or pathway level may emerge, even when the analysis of individual SNPs failed to identify any significant association.

In this context, the measure that summarizes the association between multiple SNPs and the trait of interest into a single statistic is a crucial step that raises several statistical issues. Among them, the number of SNPs considered and the impact of the possible Linkage Disequilibrium (LD) between them are often considered [4]. The most widely used approach is the minimum p-value of all the SNPs assigned to the set of SNPs, i.e. the p-value of the most significant SNP [9]. However it focuses on the most significant SNP only, rather than using the information provided by all the SNPs simultaneously which can be view as a limitation. In addition when applied directly, it has an inflated false-positive rate as it does not account for the two statistical issues described above [10]. In order to correct for both the number of SNPs and the LD, a phenotype permutation procedure can be used [11]. But permutations are time consuming, particularly if we want to reach a sufficient level of precision on p-values. Over the years, a number of alternatives have been proposed, such as the the Fisher's statistic to combine p-values of association over a set of SNPs [12].

Here we propose an adaptation of the Rasch model as a novel statistical approach to evaluate the combined effect of multiple genetic variants. Named after Georg Rasch, the Rasch model is a mathematical framework initially proposed to analyze rating scales and evaluates a latent variable not measurable directly from a set of categorical items (eg, disability, cognition or quality of life). The Rasch model is increasingly used in many areas of application such as Psychometry, Social Sciences, Education, and Clinical Trials [13], but has yet to be applied to Genetics. We believe that the application of the Rasch model to association studies offers a solution to the joint analysis of multiple genetic markers. Through different sets of simulations, the false-positive rate and power of the proposed approach is compared to a set of existing methods. By way of illustration, we also apply it to the Alzheimer ADNI GWAS data.

Methods

Introduction to the Rasch model

Some variables can be measured directly (eg, height and weight); other variables are measured indirectly by how they manifest (eg, disability, cognitive function, quality of life). Therefore, we need a method to transform the manifestations of these "latent" variables into numbers that can be taken as measurements [14]. **Rating scales** are a means to measure latent variables by a set of items, each of which has two or more ordered response categories that are assigned sequential integer scores.

For the analysis of rating scales, the **Classical Test Theory** is usually applied, whereby the item scores are summed to give a total score. However, this simple and natural approach has two main limitations [13]. First, scoring the items with sequential integers implies equal differences at the item level (differences between each response category are assumed to be equal) and at the summed score level (a change of one point implies an equal change across the range of the scale, no matter which item is concerned by this change). Consequently, such ordinal scores cannot provide us with a stable frame of reference in terms of the distance between individuals on the ability scale. Second, when applying the Classical Test Theory, the latent trait of interest is estimated by a summed score which is actually difficult to match to each single item in order to know what an individual can actually perform: individuals with the same summed score may not be able to achieve the same item task. To establish a reliable rating scale, the information of the relative difficulties of items which is actually lost in the summed score must be taken into account.

As a main alternative to overcome theses limitations, the **Item Response Theory** assumes that the probability of a specified score of a person on an item is a function of the person's ability and the item difficulty [15]:

$$\Pr(X_{ni} = x) = f(\beta_n, \tau_{ki}),$$

where $X_{ni} = x \in \{0, 1, \ldots, m_i\}$ is an integer random variable for item i where m_i is the maximum score, β_n corresponds to the ability parameter of person n and τ_{ki} corresponds to the difficulty to obtain the score k for the item i. When the person's ability is high and the item difficulty is low, the probability of having a high score for that item increases.

The **Rasch model** constitutes a particular case of the Item Response Theory and can be viewed as applying a transformation to the total scores [16]. The Rasch transformation preserves the order of the raw scores, but the distance between individuals can be assessed, and not only the rank ordering. Second, both the item difficulty and person ability are defined on the same scale; if a person's ability is known, we can predict how that person is likely to perform on an item. The Rasch model has several forms and extensions according to the data. The simplest form is the **dichotomous Rasch model** and corresponds to the situation where items have only two response categories (0 and 1). Specifically, the probability of a correct response is modeled as a logistic function of the difference between the person and item parameter:

$$\Pr(X_{ni} = 1) = \frac{\exp(\beta_n - \tau_{1i})}{1 + \exp(\beta_n - \tau_{1i})}.$$

It assumes that when the person's ability equals the item difficulty, the probability of score 1 for item i is 0.5. The **polytomous Rasch model** is a generalization of the dichotomous Rasch model [17]. Here, we will precisely consider the **Partial Credit model** which allows different

difficulty parameters for different items [14]:

$$\Pr(X_{ni} = x) = \frac{\exp \sum_{k=0}^{x} (\beta_n - \tau_{ki})}{\sum_{j=0}^{m_i} \exp \sum_{k=0}^{j} (\beta_n - \tau_{ki})}.$$

The Rasch model is based on four assumptions: 1) in the model there is only one latent variable of interest, which is the focus of the measurement and all items tap into this latent variable; 2) the total scores over an item or a person contains sufficient information for calculation of the parameters of the model; 3) for a person, the response to different items are independent; 4) the relationship between the probability of a given score to an item i and the latent trait is described by a logistic curve. Based on these assumptions, the item difficulty parameters (τ_{ki}) can be estimated by Conditional Maximum Likelihood; then the person's ability parameters (β_n) can be estimated by Maximum Likelihood.

Application of the Rasch model to multi-marker genetic association

The Rasch model is a measurement model that has potential application in any context where the objective is to measure a trait or ability through a process in which responses to items are scored with successive integers. When dealing with bi-allelic SNPs of possible alleles *a* and *A*, a set of SNPs can be considered as a set of items of possible categories 0 (= *aa*), 1 (= *aA* or *Aa*) or 2 (= *AA*) assuming an additive effect which is a reasonable hypothesis for complex traits, and analyzed with the polytomous Rasch model in order to summarize the information into one score. It corresponds to the person's ability parameter defined previously. In summary, our appraoch takes the genotypes of a set of SNPs as entry and apply the Rasch model to calculate one **multi-marker Rasch genetic score** per subject.

Once this score is estimated for each subject, its association to a trait of interest can be assessed within classical statistical inference models according to the trait of interest (linear for quantitative traits, logistic for binary traits) with the possibility to adjust with covariates such as population stratification or gender.

Implementation with R

Several softwares and R packages are available for Rasch model analysis such as `ConQuest` (https://shop.acer.edu.au/group/CON3), `RUMM` (www.rummlab.com.au), `ltm` (cran.r-project.org/package = ltm) and `eRM` (cran.r-project.org/package = eRm). Considering its flexibility and ease of integration to a pipeline of analysis, we choose to use the `eRM` R package.

The following short R script provides the functions used to obtain the multi-marker Rasch genetic score for each subject of a dataset of interest, where '`Geno`' is a data matrix of genoptypes coded by 0, 1 and 2, with subjects in rows and markers in columns:

```
> library(eRM)
> rasch.model = PCM(Geno)
> score = person.parameter(rasch.model)$theta.table[ , 1]
```

If '`Trait`' is a binary trait disease coded by 1 for cases and 0 for controls, the association of the multi-marker Rasch genetic score to the disease can then simply be assessed with a logistic model:

```
> glm(Trait ~ score, family = "binomial")
```

If 'Trait' is a quantitative trait, the association of the multi-marker Rasch genetic score to the disease can then simply be assessed with a linear model:

```
> lm(Trait ~ score)
```

Simulations

The performances of our Rasch-based multi-marker genetic association test are first evaluated in term of **false-positive rate** and **power** based on simulations over three scenarios of dependence between SNPs and varying levels of association. For each scenario, we consider:

- a binary disease trait (500 cases and 500 controls) of prevalence $K_p = 0.05$.
- a set of 24 SNPs including 12 disease susceptibility loci (DSL) simulated with relative risks ranging from 1 (no association) to 2 (strong association).

This simulation framework detailed hereafter follows principles widely used previously [18–22].

Scenario 1: SNPs are independent. The simulation model for one SNP is based on the Wright's model [23] applied to a bi-allelic marker with alleles a and A having the frequencies p_a and $p_A = 1 - p_a$. p_0, p_1 and p_2 are the frequencies of genotypes aa, aA/Aa and AA defined by the Hardy-Weinberg proportions:

$$\begin{cases} p_0 = p_a^2 + Fp_a(1-p_a) \\ p_1 = 2p_a(1-p_a) - Fp_a(1-p_a) \\ p_2 = (1-p_a)^2 + Fp_a(1-p_a) \end{cases}$$

where F is the consanguinity coefficient. This coefficient can indicate a deficit ($F > 0$) or conversely an excess ($F < 0$) of heterozygous. Here, we consider $F = 0$, so that the locus is under the Hardy-Weinberg equilibrium. We then want to compute the genotype frequencies of the SNP for cases and controls p_{Di} and p_{Hi} where $i = 0$, 1 or 2 using the disease prevalence K_p, the penetrances f_0, f_1 and f_2 of the genotypes and the mode of inheritance. The main modes of inheritance can be defined by considering the relative risks $RR_{i/0} = RR_i = \frac{f_i}{f_0}$, i = 1, 2. By assuming an additive mode of inheritance ($RR_1 = \frac{RR_2+1}{2}$), and using $f_0 = K_p/(p_0 + RR_1 \times p_1 + RR_2 \times p_2)$, $f_1 = RR_1 \times f_0$, $f_2 = RR_2 \times f_0$ and the Bayes formulas, we can easily derive the desired frequencies:

$$\begin{cases} (p_{D_0}, p_{D_1}, p_{D_2}) = \left(\frac{f_0 \times p_0}{K_p}, \frac{f_1 \times p_1}{K_p}, \frac{f_2 \times p_2}{K_p} \right) \\ (p_{H_0}, p_{H_1}, p_{H_2}) = \left(\frac{(1-f_0) \times p_0}{K_p}, \frac{(1-f_1) \times p_1}{K_p}, \frac{(1-f_2) \times p_2}{K_p} \right) \end{cases}$$

The 24 SNPs are simulated independently according to this model, the 12 non-associated SNPs with a relative risk of 1 and the 12 DSLs with a relative risk ranging from 1 to 2.

Scenario 2: SNPs in moderate Linkage Disequilibrium. To account for SNPs in Linkage Disequilibrium (LD), our simulation model follows an approach based on the diplotype frequencies of real datasets. These frequencies are used as an empirical distribution of the range of possible diplotypes. First, 12 DSLs are simulated independently from the model described in **Scenario 1**. Then the remaining SNPs are completed based on a real dataset (here the chromosome 6 of the ADNI dataset described below) in order to generate one LD blocks of moderate

magnitude (0.4–0.7) around each DSL. Simulating this way leads to genetic patterns similar to those found in real data and therefore allows us to finely control the level of LD between SNPs.

Scenario 3: SNPs in strong Linkage Disequilibrium. The simulation is the same as for **Scenario 2** with the difference that we consider SNPs in strong LD (0.8–1).

Monte-Carlo estimation of false-positive rate and power. For each scenario and each level of DSL relative risk, we ran $B = 1000$ simulations in order to provide accurate Monte-Carlo estimates of false-positive rate and power. For each simulation we obtain a p-value of association of the set of SNPs simulated by applying our Rasch-based multi-marker association test. The false-positive rate is estimated by $\Pr_{H_0}(p\text{-value} \leq \alpha)$ and the power is estimated by $\Pr_{H_1}(p\text{-value} \leqslant \alpha)$, with α the significance level usually set to 5%. Consequently in our simulations, by placing ourselves under the null hypothesis H_0 of no association ($RR_2 = 1$), then under the alternative hypothesis H_1 of association ($RR_2 > 1$), we can respectively estimate both false-positive rate and power of our method by considering the same quantity:

$$\frac{\sharp(p\text{−value}_i \leqslant \alpha, i = 1, ..., B)}{B},$$

where $\sharp()$ represents the number of p-values inferior or equal to α.

Comparison to existing methods. We compared the performances of our Rasch-based multi-marker association test to three existing methods:

- `minP` [9] is the simplest and most naive method. It considers the most significant p-value of the set of SNPs considered as the p-value of the set. This method is obviously biased since it does not take the multiple-testing and the dependence of tests into account. It is used here as a negative control and also because it is nevertheless the most widely used approach in practice.
- `GATES` [24] is a multi-marker association test using an extended Simes procedure to apply on each SNP. The p-values computed by a standard linear trend test of association on each SNP are combined with the control of correlation structure: significant p-values in high LD count less than significant p-values of independent SNPs.
- `Fisher` [12] is the well-known Fisher's combination of p-values. For m SNPs, the multi-marker test statistic is given by $T = -2\sum_{i=1}^{m} \ln(p_i)$ which has a chi-square distribution with $2m$ degrees of freedom under the null hypothesis when the m tests are independent. An adjustment to dependent tests is also available and used here [25].
- `SKAT` [26] is SNP-set Kernel Association Test. It aggregates individual test score statistics of SNPs in a set and efficiently computes the set-level p-value. It performs multiple regression of a phenotype on all variants with Davies method while adjusting for covariants for counting account for population stratification and upweights rare variants.

Application to the Alzheimer ADNI GWAS data

Alzheimer's disease (AD) is the most common neurodegenerative disorder and affects more than 35 million people worldwide. It is characterized by brain atrophy reflecting neuronal and synaptic loss and the presence of amyloid plaques and neurofibrillary tangles, leading to a progressive deterioration of cognitive functions involving memory, reason, judgment and orientation [27]. AD pathogenic mechanisms are still unclear and the disease remains a condition without cure. According to age at onset, two main types of AD are differentiated: Early-Onset AD (EOAD, appears generally before the age of 65, less than 10% of the AD population and clear genetic determinants with mutations found in the *APP*, *PSEN1* and *PSEN2* genes) and

Late-Onset AD (LOAD, more than 90% of the AD population, appears generally after the age of 65 and has a complex etiology based on genetic and environmental factors) [28].

In recent years, several Genome-Wide Association Studies (GWAS) were performed to detect genetic loci associated with LOAD [29–31]. These studies support the hypothesis that *APOE* is a major susceptibility gene for LOAD [32]. In addition to *APOE*, markers within several other genes gave replicated evidence of association with LOAD [33]. The identification of these genes improves our knowledge of AD. For instance, *CR1* has been demonstrated to be able to produce an AD up-regulated protein [34]. Although these new loci have been found, some problems ramain unsolved. First, to date none of these loci has proven accurate or sensitive enough to serve as biomarker. Second, the replication of results is a tedious task in GWAS. To push the boundaries of current knowledge on AD, further studies about GWAS and statistical models are still necessary.

By way of illustration, we applied our Rasch-based multi-marker association test to the genes of the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) [31]. The study population is made up of 359 cases and 226 controls, genotyped with an Illumina Human 610-Quad (= 620901 SNPs). A standard quality control process based on minor allele frequency, Hardy-Weinberg equilibrium, missingness and relatedness excluded 31 cases, 49 controls and 82071 SNPs [35]. The dataset was also reduced with a minimal loss of information by pruning with `Plink` (window size = 50 SNPs, shift = of 5 SNPs at each step and threshold correlation coefficient of 0.2) [36]. Missing genotypes were imputed with weighted k-Nearest-Neighbors method [37]. SNPs are considered attached to a gene if they are located within a distance of 20 kb around it. The curated dataset to analyze comprises 16514 genes. For each gene and each subject, a Rasch-based multi-marker genetic score is computed, and the association of this score to the disease is evaluated by a logistic regression model.

The top genes identified by the Rasch analysis were integrated into a hypothetic signalling network. Protein-protein interaction data and functional findings were extracted from QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity), manually analysed and supplemented by literature curation.

Results

Simulations

False-positive rate and power for `Rasch`, `minP`, `GATES`, `Fisher` and `SKAT` across the three scenarios are given Fig 1. The first observation is that `minP` has a strongly inflated false-positive rate, far above the expected 5% level and decreasing with the level of LD (0.691 for **Scenario 1**, 0.285 for **Scenario 2** and 0.145 for **Scenario 3**). This observation was actually expected knowing the drawbacks of the `minP` method, and validates our simulations. On the other hand, `Rasch`, `GATES`, `Fisher` and `SKAT` have a correct control of the false-positive rate to 5% (Fig 1a,1b,1c) and a power that increases toward 100% with an increasing level of association to the disease (Fig 1d,1e,1f). In term of false-positive rate, it is worthy to mention that on **Scenario 3** (Fig 1c) `Rasch` is the closest to the 5% level (estimated to 0.043) whereas `GATES` and `Fisher` are more conservative (estimated to 0.039 and 0.034 respectively) and `SKAT` is more inflated. In term of power, `Rasch` has the best performances on independent SNPs followed by `Fisher` (**Scenario 1**, Fig 1d). Both methods have similar good performances when applied on SNPs with moderate and strong LD (**Scenarios 2,3**, Fig 1e,1f). The performance of `GATES` is better compared to `SKAT` on independent SNPs but is limited on LD block simulation. (Fig 1d,1e,1f).

Table 1. Annotation of the 20 top genes resulting from the application of the Rasch-based multi-marker association test to the genes of the Alzheimer ADNI GWAS dataset.

Chromosome	Gene symbol	*p*-value	Full name	Location	Function
19	APOE	2.30e-08	Apolipoprotein E	Extracellular space	Transporter
7	ZNF398	9.71e-06	Zinc finger protein 398	Nucleus	Transcription regulator
15	AEN	1.28e-05	Apoptosis enhancing nuclease	Nucleus	Enzyme
7	ZNF425	1.37e-04	Zinc finger protein 425	Other	Other
5	ADAMTS12	1.59e-04	ADAM metallopeptidase with thrombospondin	Extracellular space	Peptidase
1	PSMA5	2.41e-04	Proteasome (prosome, macropain) subunit	Cytoplasm	Peptidase
13	FAM124A	2.85e-04	Family with sequence similarity 124A	Other	Other
9	FXN	4.29e-04	Frataxin	Cytoplasm	Kinase
11	NTM	4.80e-04	Neurotrimin	Plasma Membrane	Other
5	LARP1	5.01e-04	La ribonucleoprotein domain family	Cytoplasm	Translation regulator
1	WDTC1	5.39e-04	WD and tetratricopeptide repeats 1	Other	Other
11	EPS8L2	5.94e-04	EPS8-like 2	Other	Other
3	KBTBD12	6.27e-04	Kelch repeat and BTB domain containing 12	Other	Other
7	FAM188B	6.36e-04	Family with sequence similarity 188	Other	Other
17	OR3A3	6.94e-04	Olfactory receptor, family 3	Plasma membrane	G-protein coupled receptor
2	BZW1	7.17e-04	Basic leucine zipper and W2 domains 1	Cytoplasm	Translation regulator
23	TMEM187	7.62e-04	Transmembrane protein 187	Cytoplasm	Other
15	SEMA7A	7.87e-04	Semaphorin 7A, GPI membrane anchor	Plasma membrane	Transmembrane receptor
7	VKORC1L1	8.23e-04	Vitamin K epoxide reductase complex	Cytoplasm	Enzyme
19	COL5A3	9.80e-04	Collagen, type V, alpha 3	Extracellular space	Other

doi:10.1371/journal.pone.0138223.t001

deviate from the QQ-line (Fig 2): *ZNF398* ($p = 9.71e^{-6}$) and *AEN* ($p = 1.27e^{-5}$). *AEN* encodes an enhancing apoptosis nuclease, a process that takes part to the neuronal loss observed in AD. We unfortunately did not find any indication about the possible functional implication of *ZNF398* in AD.

As we noticed a slight deviation from the QQ-line at 10^{-3} (Fig 2), we also investigated the other 17 top genes. Several of them could be functionally linked to AD:

- *PSMA5* is a proteasome subunit involved in the apoptosis process that takes part in the neuronal loss observed in AD [39]. *PSMA5* was also found to interact directly with the AD associated *PSEN1* gene [40].
- *FXN* encodes the frataxin mitochondrial protein which functions in regulating mitochondrial iron transport and respiration. Frataxin deficiency leads to mitochondrial dysfunction and oxidative damage that are at the origin of numerous neurodegenerative diseases like Friedreich ataxia, Parkinson and AD [41]. Interestingly, another top gene *VKORC1L1* is also involved in regulation of oxidative stress and mediates vitamin K-dependent intracellular antioxidant function [42]. Remarkably, blood level of vitamin K in *APOE4* carriers is lower than in persons with other *APOE* genotypes implying hypothetical link of vitamin K deficiency to pathogenesis of AD [43, 44].
- Alzheimer's disease is sometimes named 'type 3 diabetes' due to twice more frequent occurrence in diabetic patients [45, 46]. Two top genes from our list (*COL5A3* and *WDTC1*) were identified as potent modulators of insulin signalling [47, 48]. Noteworthy,

vitamin K-dependent modification of osteocalcin was also shown to affect glucose homeostasis [49].

- *NTM* encodes a neural cell adhesion molecule that modulates neurite outgrowth and adhesion via a homophilic mechanism [50]. Some data indicates that *NTM* might directly bind to amyloid beta [51]. It has been associated to intelligence in a family-based association study [52] and lies at locus 11q25 which has been associated with AD [53]
- *SEMA7A* belongs to the semaphorins family involved in neuronal processes. Semaphorins and their downstream signaling components regulate synaptic physiology and neuronal excitability in the mature hippocampus, and these proteins are also implicated in a number of developmental, psychiatric, and neurodegenerative disorders [54]. Remarkably, *SEMA7A* not only enhances axon growth via beta1-integrin, but equally processes immune-modulatory activity and regulates endothelial functions [55, 56]. As well, another top gene (*ADAMTS12*) is also implicated in control of immune responses and angiogenesis, deregulated in course of Alzheimer's disease [57, 58].
- Finally, *LARP1* protein associates with the mTOR complex 1 (mTORC1) regulating global protein synthesis. Functional importance of mTOR signalling has been experimentally confirmed in Alzheimer's disease, and therapeutic targeting of this signalling module is considered as a promising strategy for developing neuro-protective treatments [59–61].

We also performed a formalized network analysis based on these top genes with the Ingenuity Pathway Analysis. The resulting network is given Fig 3 and seems to highlight the metabolism of cholesterol that plays a key role in AD pathogenesis [62–64]. Nine of the 20 top genes are connected in this network (in orange). Remarkably, most of them can be functionally linked to AD. For example, integrin *ITGB1* mediates effect of *SEMA7A* on axon growth [65]. The integrins are modulated by *CASR* gene that forms a functional complex with metabotropic glutamate receptor *GRM5* [66, 67]. It was recently shown that *GRM5* is a co-receptor for cytotoxic Aβ oligomers bound to prion *PRNP* protein [68].

Discussion

With the recent improvement of high-throughput genotyping technologies, the use of Genome-Wide Association Studies has become widespread in genetic research. However, the high dimension of the genetic data, the simultaneous testing of many markers and the necessity to account for the complex genetic structure of human populations are, among others, tricky issues that have raised doubts about the relevance of these studies' findings. The development of methods in Statistical Genetics is therefore very important to ensure that such studies are correctly conducted and to provide a proper interpretation of their findings, and this research has involved scientists from many disciplines. In this context, applying the Rasch model initially developed for psychometric data to the analysis of genetic data can be viewed as a new link between two areas of research that was not obvious before. Our novel statistical approach may be useful to complement at the gene or pathway level, the findings of significant associations made at the single SNP level.

Based on simulations, it showed in different situations good performances in terms of false-positive rate and power compared to other popular methods (`minP`, `GATES`, `Fisher` and `SKAT`). We noticed that the benefits of Rasch in terms of power were more important when applied to independent SNPs which is coherent with one of the assumptions of the model that the response to different items are independent. As this loss of power is observed for all the methods when the level of dependence between the SNPs (Linkage Disequilibrium) increases,

treat alcohol-dependence [69]. In combination with baclofen, acamprosate has recently been shown to be effective over a range of preclinical AD models [70], and has demonstrated promising results in phase 2a clinical trial for AD [71].

Through this study, we encountered three limitations for the application of the Rasch model. First, it works on complete data without missing values. However missing values are a common problem in most scientific research domains as they can arise from different sources such as mishandling of samples, low signal-to-noise ratio, measurement error, non-response or deleted aberrant value. Consequently the application of the Rasch model requires preliminary imputation of missing values. This imputation is a general and separate scientific topic that has been thoroughly discussed to date [72–76]. Second, in some particular cases the estimation of the Rasch model with the `eRM` R package does not converge and consequently does not provide any results. It happened for instance to 9 genes over the 16514 genes analyzed in the ADNI GWAS data and the reasons of that problem were not clear to us. Finally applying a Rasch model necessitates accessing individual level genetic data. But often, only summary statistics are available for published GWAS. This is a real limitation for most of the existing multi-marker methods in order to correctly account for gene size and LD, although some authors have found a solution in using the genotype data from a reference panel such as the 1000 Genomes or the HapMap projects [77–79] which is not applicable here.

The application of the Rasch model also opens two opportunities that were not yet considered here. The analysis of multiple markers is not limited to the gene level, and the Rasch-based multi-marker genetic association test could also be applied to the analysis whole pathways. In addition, this genetic score could also be used as a predictor of the disease for the supervised classification of cases versus controls. The Rasch model is also suitable to the inclusion of rare variants, as most rare variants analyses focus on gene level test by collapsing the effects of all rare SNPs in a gene into a single test of association [4]. These applications deserve further investigation.

From a broader point of view, given the urgent need to understand how the thousands of loci that have been identified in genome-wide association studies contribute to the genetic basis complex traits, the application of multi-marker methods at the gene or pathway level becomes an increasingly important approach for secondary analysis of GWAS data [80–82]. Main recognized benefits include the incorporation of biological knowledge, the reduction in multiple-testing and the consideration of SNPs with modest effects. But this type of analysis has also clear limitations [4]. For instance determining whether a particular SNP is part of, or regulates a gene is a thorny problem. In addition, by focusing on SNPs that can be assigned to genes, analyzing GWAS data at the gene level also misses many disease associated SNPs that cannot be linked to genes (such as SNPs in gene deserts for instance). In that case, the delimitation of genomic regions made of contiguous SNPs and associated as a whole, should also complement our understanding of the genetic of complex traits [20].

Acknowledgments

We thank Linda Belo and Julien Laffaire for their critical reviewing of the manuscript. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI-PI is Michael W. Weiner, MD(email: Michael.Weiner@ucsf.edu). A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgment_List.pdf. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from

the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego.

Author Contributions

Conceived and designed the experiments: WW MG IC D. Cohen. Performed the experiments: WW MG JM JB. Analyzed the data: WW MG JM D. Commenges SN IC D. Cohen. Wrote the paper: WW MG. Performed the Systems Biology analysis: SN.

References

1. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21. doi: 10.1038/456018a PMID: 18987709
2. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, et al. (2008) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature genetics 41: 25–34. doi: 10.1038/ng.287 PMID: 19079261
3. Lehne B, Lewis CM, Schlitt T (2011) From snps to genes: disease association at the gene level. PloS one 6: e20133. doi: 10.1371/journal.pone.0020133 PMID: 21738570
4. Fridley BL, Biernacka JM (2011) Gene set analysis of snp data: benefits, challenges, and future directions. European Journal of Human Genetics 19: 837–843. doi: 10.1038/ejhg.2011.57 PMID: 21487444
5. Jorgenson E, Witte JS (2006) A gene-centric approach to genome-wide association studies. Nature Reviews Genetics 7: 885–891. doi: 10.1038/nrg1962 PMID: 17047687
6. Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. The American Journal of Human Genetics 75: 353–362. doi: 10.1086/423901 PMID: 15272419
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102: 15545–15550. doi: 10.1073/pnas.0506580102 PMID: 16199517
8. Bouaziz M, Jeanmougin M, Guedj M (2012) Multiple testing in large-scale genetic studies. In: Data Production and Analysis in Population Genomics, Springer. pp. 213–233.
9. Torkamani A, Topol EJ, Schork NJ (2008) Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92: 265–272. doi: 10.1016/j.ygeno.2008.07.011 PMID: 18722519
10. Hong MG, Pawitan Y, Magnusson PK, Prince JA (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Human Genetics 126: 289–301. doi: 10.1007/s00439-009-0676-z PMID: 19408013
11. Evangelou M, Rendon A, Ouwehand WH, Wernisch L, Dudbridge F (2012) Comparison of methods for competitive tests of pathway analysis. PloS one 7: e41018. doi: 10.1371/journal.pone.0041018 PMID: 22859961
12. Whitlock M (2005) Combining probability from independent tests: the weighted z-method is superior to fisher's approach. Journal of evolutionary biology 18: 1368–1373. doi: 10.1111/j.1420-9101.2005.00917.x PMID: 16135132
13. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ (2007) Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. The Lancet Neurology 6: 1094–1105. doi: 10.1016/S1474-4422(07)70290-9 PMID: 18031706

14. Masters GN (1982) A rasch model for partial credit scoring. Psychometrika 47: 149–174. doi: 10.1007/BF02296272
15. Woods A, Baker R (1985) Item response theory. Language Testing 2: 117–140. doi: 10.1177/026553228500200202
16. Rasch G (1960) Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.
17. Andrich D, Sheridan B, Lyne A, Luo G (2000) Rumm: A windows-based item analysis program employing rasch unidimensional measurement models. Perth, Australia: Murdoch University.
18. Bouaziz M, Ambroise C, Gued M (2011) Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. PloS one 6(12): e28845. doi: 10.1371/journal.pone.0028845 PMID: 22216125
19. Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Human heredity 56: 18–31. doi: 10.1159/000073729 PMID: 14614235
20. Guedj M, Robelin D, Hoebeke M, Lamarine M, Wojcik J, et al. (2006) Detecting local high-scoring segments: A first-stage approach for genome-wide association studies. Statistical applications in genetics and molecular biology 5. doi: 10.2202/1544-6115.1192 PMID: 17049033
21. Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. The american journal of human genetics 80: 353–360. doi: 10.1086/511312 PMID: 17236140
22. Pan W (2009) Asymptotic tests of association with multiple snps in linkage disequilibrium. Genetic epidemiology 33: 497–507. doi: 10.1002/gepi.20402 PMID: 19170135
23. Wright BD, Masters GN (1982) Rating Scale Analysis. Chicago: MESA Press.
24. Li MX, Gui HS, Kwan JS, Sham PC (2011) Gates: a rapid and powerful gene-based association test using extended simes procedure. The American Journal of Human Genetics 88: 283–293. doi: 10.1016/j.ajhg.2011.01.019 PMID: 21397060
25. Moskvina V, O'Dushlaine C, Purcell S, Craddock N, Holmans P, et al. (2011) Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. Genetic epidemiology 35: 861–866. doi: 10.1002/gepi.20636 PMID: 22006681
26. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82–93. doi: 10.1016/j.ajhg.2011.05.029 PMID: 21737059
27. Salek RM, Xia J, Innes A, Sweatman BC, Adalbert R, et al. (2010) A metabolomic study of the crnd8 transgenic mouse model of alzheimer's disease. Neurochemistry international 56: 937–947. doi: 10.1016/j.neuint.2010.04.001 PMID: 20398713
28. Rogaeva E (2002) The solved and unsolved mysteries of the genetics of early-onset alzheimer's disease. Neuromolecular medicine 2: 1–10. doi: 10.1385/NMM:2:1:01 PMID: 12230301
29. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, et al. (2009) Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. Nature genetics 41: 1088–1093. doi: 10.1038/ng.440 PMID: 19734902
30. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, et al. (2010) Genome-wide analysis of genetic loci associated with alzheimer disease. Jama 303: 1832–1840. doi: 10.1001/jama.2010.574 PMID: 20460622
31. Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, et al. (2009) Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for alzheimer's disease. PloS one 4: e6501. doi: 10.1371/journal.pone.0006501 PMID: 19668339
32. Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, et al. (2007) A high-density whole-genome association study reveals that apoe is the major susceptibility gene for sporadic late-onset alzheimer's disease. Journal of Clinical Psychiatry 68: 613. doi: 10.4088/JCP.v68n0419 PMID: 17474819
33. Lee JH, Cheng R, Barral S, Reitz C, Medrano M, et al. (2011) Identification of novel loci for alzheimer disease and replication of clu, picalm, and bin1 in caribbean hispanic individuals. Archives of Neurology 68: 320–328. doi: 10.1001/archneurol.2010.292 PMID: 21059989
34. Corneveaux JJ, Myers AJ, Allen AN, Pruzin JJ, Ramirez M, et al. (2010) Association of cr1, clu, and picalm with alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. Human molecular genetics: ddq221.
35. Weale ME (2010) Quality control for genome-wide association studies. In: Genetic Variation, Springer. pp. 341–372.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81: 559–575. doi: 10.1086/519795 PMID: 17701901

37. Schwender H (2012) Imputing missing genotypes with weighted k nearest neighbors. Journal of Toxicology and Environmental Health, Part A 75: 438–446. doi: 10.1080/15287394.2012.674910

38. Corder E, Saunders A, Strittmatter W, Schmechel D, Gaskell P, et al. (1993) Gene dose of apolipoprotein e type 4 allele and the risk of alzheimer's disease in late onset families. Science 261: 921–923. doi: 10.1126/science.8346443 PMID: 8346443

39. Wu Y, Deng Y, Zhang S, Luo Y, Cai F, et al. (2015) Amyloid-β precursor protein facilitates the regulator of calcineurin 1-mediated apoptosis by downregulating proteasome subunit α type-5 and proteasome subunit β type-7. Neurobiology of aging 36: 169–177. doi: 10.1016/j.neurobiolaging.2014.07.029 PMID: 25194880

40. Van Gassen G, De Jonghe C, Pype S, Van Criekinge W, Julliams A, et al. (1999) Alzheimer's disease associated presenilin 1 interacts with hc5 and zeta, subunits of the catalytic 20s proteasome. Neurobiology of disease 6: 376–391. doi: 10.1006/nbdi.1999.0265 PMID: 10527805

41. Gomes CM, Santos R (2013) Neurodegeneration in friedreich's ataxia: From defective frataxin to oxidative stress. Oxidative medicine and cellular longevity 2013. doi: 10.1155/2013/487534

42. Westhofen P, Watzka M, Marinova M, Hass M, Kirfel G, et al. (2011) Human vitamin k 2, 3-epoxide reductase complex subunit 1-like 1 (vkorc1l1) mediates vitamin k-dependent intracellular antioxidant function. Journal of Biological Chemistry 286: 15085–15094. doi: 10.1074/jbc.M110.210971 PMID: 21367861

43. Allison A (2001) The possible role of vitamin k deficiency in the pathogenesis of alzheimer's disease and in augmenting brain damage associated with cardiovascular disease. Medical hypotheses 57: 151–155. doi: 10.1054/mehy.2001.1307 PMID: 11461163

44. Huy PDQ, Yu YC, Ngo ST, Thao TV, Chen Cp, et al. (2013) In silico and in vitro characterization of anti-amyloidogenic activity of vitamin k3 analogues for alzheimer's disease. Biochimica et Biophysica Acta (BBA)-General Subjects 1830: 2960–2969. doi: 10.1016/j.bbagen.2012.12.026

45. Chen Z, Zhong C (2013) Decoding alzheimer's disease from perturbed cerebral glucose metabolism: implications for diagnostic and therapeutic strategies. Progress in neurobiology 108: 21–43. doi: 10.1016/j.pneurobio.2013.06.004 PMID: 23850509

46. Blázquez E, Velázquez E, Hurtado-Carneiro V, Ruiz-Albusac JM (2014) Insulin in the brain: its pathophysiological implications for states related with central insulin resistance, type 2 diabetes and alzheimer's disease. Frontiers in endocrinology 5. doi: 10.3389/fendo.2014.00161 PMID: 25346723

47. Suh JM, Zeve D, McKay R, Seo J, Salo Z, et al. (2007) Adipose is a conserved dosage-sensitive antiobesity gene. Cell metabolism 6: 195–207. doi: 10.1016/j.cmet.2007.08.001 PMID: 17767906

48. Huang G, Ge G, Wang D, Gopalakrishnan B, Butz DH, et al. (2011) α3 (v) collagen is critical for glucose homeostasis in mice due to effects in pancreatic islets and peripheral tissues. The Journal of clinical investigation 121: 769–783. doi: 10.1172/JCI45096 PMID: 21293061

49. Shiba S, Ikeda K, Azuma K, Hasegawa T, Amizuka N, et al. (2014) γ-glutamyl carboxylase in osteoblasts regulates glucose metabolism in mice. Biochemical and biophysical research communications 453: 350–355. doi: 10.1016/j.bbrc.2014.09.091 PMID: 25264202

50. Gil OD, Zanazzi G, Struyk AF, Salzer JL (1998) Neurotrimin mediates bifunctional effects on neurite outgrowth via homophilic and heterophilic interactions. The Journal of neuroscience 18: 9312–9325. PMID: 9801370

51. Oláh J, Vincze O, Virók D, Simon D, Bozsó Z, et al. (2011) Interactions of pathological hallmark proteins: tubulin polymerization promoting protein/p25, beta-amyloid, and alpha-synuclein. The Journal of Biological Chemistry 286: 34088–34100. doi: 10.1074/jbc.M111.243907 PMID: 21832049

52. Pan Y, Wang KS, Aragam N (2011) Ntm and nr3c2 polymorphisms influencing intelligence: family-based association studies. Progress in Neuro-Psychopharmacology and Biological Psychiatry 35: 154–160. doi: 10.1016/j.pnpbp.2010.10.016 PMID: 21036197

53. Blacker D, Bertram L, Saunders AJ, Moscarillo TJ, Albert MS, et al. (2003) Results of a high-resolution genome screen of 437 alzheimer's disease families. Human Molecular Genetics 12: 23–32. doi: 10.1093/hmg/ddg007 PMID: 12490529

54. Pasterkamp RJ, Giger RJ (2009) Semaphorin function in neural plasticity and disease. Current opinion in neurobiology 19: 263–274. doi: 10.1016/j.conb.2009.06.001 PMID: 19541473

55. Suzuki K, Okuno T, Yamamoto M, Pasterkamp RJ, Takegahara N, et al. (2007) Semaphorin 7a initiates t-cell-mediated inflammatory responses through &agr; 1&bgr; 1 integrin. Nature 446: 680–684. doi: 10.1038/nature05652 PMID: 17377534

56. Morote-Garcia JC, Napiwotzky D, Köhler D, Rosenberger P (2012) Endothelial semaphorin 7a promotes neutrophil migration during hypoxia. Proceedings of the National Academy of Sciences 109: 14146–14151. doi: 10.1073/pnas.1202165109

57. El Hour M, Moncada-Pazos A, Blacher S, Masset A, Cal S, et al. (2010) Higher sensitivity of adamts12-deficient mice to tumor growth and angiogenesis. Oncogene 29: 3025–3032. doi: 10.1038/onc.2010.49 PMID: 20208563

58. Moncada-Pazos A, Obaya AJ, Llamazares M, Heljasvaara R, Suárez MF, et al. (2012) Adamts-12 metalloprotease is necessary for normal inflammatory response. Journal of Biological Chemistry 287: 39554–39563. doi: 10.1074/jbc.M112.408625 PMID: 23019333

59. Caccamo A, De Pinto V, Messina A, Branca C, Oddo S (2014) Genetic reduction of mammalian target of rapamycin ameliorates alzheimer's disease-like cognitive and pathological deficits by restoring hippocampal gene expression signature. The Journal of Neuroscience 34: 7988–7998. doi: 10.1523/JNEUROSCI.0777-14.2014 PMID: 24899720

60. Maiese K (2014) Taking aim at alzheimer's disease through the mammalian target of rapamycin. Annals of medicine 46: 587–596. doi: 10.3109/07853890.2014.941921 PMID: 25105207

61. Tcherkezian J, Cargnello M, Romeo Y, Huttlin EL, Lavoie G, et al. (2014) Proteomic analysis of cap-dependent translation identifies larp1 as a key regulator of 5 top mrna translation. Genes & development 28: 357–371. doi: 10.1101/gad.231407.113

62. Crameri A, Biondi E, Kuehnle K, Lütjohann D, Thelen KM, et al. (2006) The role of seladin-1/dhcr24 in cholesterol biosynthesis, app processing and aβ generation in vivo. The EMBO journal 25: 432–443. doi: 10.1038/sj.emboj.7600938 PMID: 16407971

63. Stefani M, Liguri G (2009) Cholesterol in alzheimer's disease: unresolved questions. Current Alzheimer Research 6: 15–29. doi: 10.2174/156720509787313899 PMID: 19199871

64. Sun JH, Yu JT, Tan L (2014) The role of cholesterol metabolism in alzheimer's disease. Molecular neurobiology: 1–19.

65. Pasterkamp RJ, Peschon JJ, Spriggs MK, Kolodkin AL (2003) Semaphorin 7a promotes axon outgrowth through integrins and mapks. Nature 424: 398–405. doi: 10.1038/nature01790 PMID: 12879062

66. Gama L, Wilt SG, Breitwieser GE (2001) Heterodimerization of calcium sensing receptors with metabotropic glutamate receptors in neurons. Journal of Biological Chemistry 276: 39053–39059. doi: 10.1074/jbc.M105662200 PMID: 11489900

67. Tharmalingam S, Daulat AM, Antflick JE, Ahmed SM, Nemeth EF, et al. (2011) Calcium-sensing receptor modulates cell adhesion and migration via integrins. Journal of Biological Chemistry 286: 40922–40933. doi: 10.1074/jbc.M111.265454 PMID: 21969374

68. Um JW, Kaufman AC, Kostylev M, Heiss JK, Stagi M, et al. (2013) Metabotropic glutamate receptor 5 is a coreceptor for alzheimer aβ oligomer bound to cellular prion protein. Neuron 79: 887–902. doi: 10.1016/j.neuron.2013.06.036 PMID: 24012003

69. Mason BJ (2001) Treatment of alcohol-dependent outpatients with acamprosate: a clinical review. Journal of Clinical Psychiatry 62: 42–48. PMID: 11584875

70. Chumakov I, Nabirotchkin S, Cholet N, Milet A, Boucard A, et al. (2015) Combining two repurposed drugs as a promising approach for alzheimer's disease therapy. Scientific reports 5. doi: 10.1038/srep07608 PMID: 25566747

71. Scart-Gres C, Haddad R, Foucquier J, Guedj M, Hugon J, et al. (2014) First evidence of pxt00864 effect in the treatment of mild ad: results on 30 patients from the pleodial study. CTAD abstracts, The Journal of Prevention of Alzheimer's Disease,Volume 1, Number 3, 232–233.

72. Rubin DB (2004) Multiple imputation for nonresponse in surveys, volume 81. John Wiley & Sons.

73. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17: 520–525. doi: 10.1093/bioinformatics/17.6.520 PMID: 11395428

74. Oba S, Sato Ma, Takemasa I, Monden M, Matsubara Ki, et al. (2003) A bayesian missing value estimation method for gene expression profile data. Bioinformatics 19: 2088–2096. doi: 10.1093/bioinformatics/btg287 PMID: 14594714

75. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics 81: 1084–1097. doi: 10.1086/521987 PMID: 17924348

76. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics 5: e1000529. doi: 10.1371/journal.pgen.1000529 PMID: 19543373

77. Consortium GP, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi: 10.1038/nature09534

78. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, et al. (2010) A versatile gene-based test for genome-wide association studies. The American Journal of Human Genetics 87: 139–145. doi: 10.1016/j.ajhg.2010.06.009 PMID: 20598278

79. Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, et al. (2014) A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. Genetic epidemiology 38: 661–670. doi: 10.1002/gepi.21853 PMID: 25371288

80. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, et al. (2007) A genomic pathway approach to a complex disease: axon guidance and parkinson disease. PLoS genetics 3: e98. doi: 10.1371/journal.pgen.0030098 PMID: 17571925

81. Eleftherohorinou H, Wright V, Hoggart C, Hartikainen AL, Jarvelin MR, et al. (2009) Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases. PloS one 4: e8068. doi: 10.1371/journal.pone.0008068 PMID: 19956648

82. Carbonetto P, Stephens M (2013) Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for il-2 signaling genes in type 1 diabetes, and cytokine signaling genes in crohn's disease. PLoS genetics 9: e1003770. doi: 10.1371/journal.pgen.1003770 PMID: 24098138

Chapter 4

Item Response Theory Based Mutual Information Test

4.1. Introduction

In the previous two chapters, we applied the Rasch model to different categorical data: rating scale and GWAS data. Rasch model has demonstrated its capacity on the item evaluation and person ability estimation. Still, the utilization of Rasch model and Item Response Theory is not limited to the direct analysis of the parameters that they produced through fitting the data. These parameters could be employed in other statistical models to provide a more precise data estimation. For instance, an item which is more dependent on the latent trait has higher possibility to correspond to the severity change of disease and thus to be a sensitive marker for this population. Mutual information in the field of information theory provides a general dependency estimation by quantifying the dependence between the joint distribution of two variables. The probability density distribution of the items can be estimated by Item Response Theory. In this chapter, we proposed a new method of items evaluation with the mutual information based on the Item Response Theory. The aim was to select the sensitive items in ADAS-cog, the most used cognitive functioning measures in Alzheimer's Disease, in order to better serve the disease progress evaluation.

An article resulted by this study titled as "Selection of items as sensitive clinical markers for MCI population from the ADAS-Cog with the IRT-based mutual information" is in preparation.

4.2. Method and Results

In this study, we calculated the IRT-based mutual information of each item in ADAS-cog using the baseline data of MCI patients in the ADNI study. The two-parameter IRT polytomous model Graded Response Model was applied to estimate the probability density function. This model was implemented with the *R* package *ltm.*

To verify if the mutual information reveals the sensitivity of items, we compared it with other IRT-based statistics: the Fisher information and discrimination. The top items selected by the three methods were similar, but globally the mutual information of items is better correlated to the severity change in the two years follow-up data compared to other methods.

To reduce the variability and enhance the reliability, it is of interest to select the more sensitive sub-items in the scale. Items Word Recall, Word Recognition and Delayed Word Recall were found to have higher mutual information in the ADAS-cog. Their composite score showed a higher rate of change compared to other composite scores of the ADAS-cog subscales. These studies may also indicate that the high mutual information items could be sensitive markers for the early stage of the disease such as MCI.

4.3. Manuscript

Selection of ADAS-Cog items for MCI population by IRT-based mutual information

Wenjia Wang[1,2,*], Mickaël Guedj[2], Helene Jacqmin-Gadda[1], Cecile Proust-Lima[1] and Daniel Commenges[1], the Alzheimer's Disease Neuroimaging Initiative [¶]

[1]Inserm U1219, Université de Bordeaux, Bordeaux, Aquitaine, France.

[2]Pharnext, Issy-Les-Moulineaux, France.

¶Membership of the Alzheimer's Disease Neuroimaging Initiative can be found in the Acknowledgments section.

*Corresponding author: wenjia.wang@pharnext.com

Abstract

Although the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog) is the most used cognitive functioning measures in Alzheimer's Disease, studies showed that its sensitivity on measuring disease progression in clinical trials is limited, especially in MCI patients. It is needed to select sensitive items in ADAS-cog for the purpose of helping to identify and to treat patients in the early stage of the disease. In this study, we proposed a new method to evaluate the sensitivity of items with the Item Response Theory (IRT) based mutual information. In comparison with other IRT-based statistics, the mutual information of items better corresponds to items' rate of change in follow-up data. The composite score of items with high mutual information shows a higher rate of change compared to other subscales of ADAS-cog. In conclusion, this IRT-based mutual information could be a useful statistic in the sensitive item selection in measures.

Introduction

Alzheimer's Disease (AD) is the most common neurodegenerative disorder and there is no available therapeutic treatment up to date. Recent Researches suggests that AD begins years

before the development of symptoms (Skinner et al., 2012). Slowly, patients typically perform more poorly on tests of memory, language, executive function and visuospatial ability as part of disease progression (Koppel, 2005). Early detection of cognitive changes in the preclinical stage or Mild Cognitive Impairment (MCI) is crucial since many clinical trials are targeting modification underlying disease pathology rather than ameliorate symptoms for AD.

Global cognitive functioning measures are essential tools for diagnosis and progression tracking of AD. Among numerous tests available for cognitive dysfunction, the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog) has been the most widely used scale in antidementia clinical trials in patients with mild and moderate AD. The original 11-item ADAS scale (ADAS-cog 11 or ADAS-classic) was developed by Rosen *et al.* in 1984 (Rosen, Mohs, and Davis 1984). Skinner *et al.* added items Delayed Word Recall and Number Cancellation on the ADAS-cog 11 and this modified version is called ADAS-cog 13 (Skinner et al., 2012).

Although some studies have demonstrated its use as an effective measure of dementia severity and progression (Ihl et al., 1999; Weyer et al., 1997), ADAS-cog still has some limitations: it did not distinguish reliably the different cognitive impairment levels (Ihl et al., 1992); the scale is also not uniformly sensitive to measuring cognitive decline in AD, especially for MCI or mild AD patients. This may be due to the inequality of the sensitivity of items to detect cognitive deficits. Some items demonstrate a ceiling effect that makes them uninformative in subjects within predementia stage (Cano et al. 2010). The inclusion of insensitive items also brings in more variability, which potentially obscured the mild deficits tracking.

Given the current state of clinical research in AD, a selection of items in ADAS-cog in MCI patients based on their sensitivity is desired. By eliminating less informative items, we could reduce variability for predementia clinical researches and the test process could also be more efficient. The latent trait that ADAS-cog tries to estimate is the cognitive disability led by the disease. The cognitive functions worsen with the disease progression. When the response to an item in the scale is more dependent to the latent trait based on the data of a certain population, it has higher possibility to correspond to the severity change of disease and thus to be a sensitive marker for this population.

The evaluation of dependency is an important issue in many problems (Karasuyama and Sugiyama, 2012; Steuer et al., 2002). Several measures quantify the dependency between observed random variables, such as the Pearson correlation coefficient and the Spearman correlation coefficient. Compared to these measures, mutual information provides a general dependency estimation by quantifying the dependence between the joint distribution of two variables and what the joint distribution would be if they were independent. High mutual information indicates a large reduction in uncertainty of one random variable due to the knowledge of the other. Mutual information has been widely applied in other statistical decision contexts finding important applications as an indicator in feature extraction (Silva and Narayanan, 2009), detection (Cooper, 2000), image registration and segmentation (Thévenaz et al., 2000), and to characterize performance limits on pattern recognition (Westover and O'Sullivan, 2008). The properties of mutual information make it possible to be extended to tests where an estimate of the continuous latent trait is desired. The application of mutual information approaches to item selection have been considered mostly in the context of computerized adaptive testing (Liu, 2005; Wang, 2013) but has not yet been utilized for item selection of clinical scales.

The mutual information of two random variables depends on their distributions. However, the distribution of the item response is unknown in most cases. In ADAS-cog, the difficulties of items vary and the categories increase indicating from low to high the severity of the disease. To estimate the mutual information between the latent trait and the items, we need to estimate the probability density function for each category from the sample.

Item Response Theory (IRT) provides a statistical framework of measurement analysis that can be used to approximate probability density functions in measurement. Recent studies have demonstrated that the application of IRT increased the precision in the cognitive assessment (Ard et al., 2013; Balsis et al., 2012). Compared to the traditional way that uses the total scores in the diagnosis and assessment of a measure, IRT considers the situation that some items may be more difficult than others and the capacity of subjects varies (Hambleton et al., 1991). IRT assumes that the probability of certain response on an item is a mathematical function of the person and item parameters. The person parameter is the estimation of the latent trait; the item parameters are statistics expressing the relationship between the outcome of items and the latent trait, and they could be used to evaluate the psychometric properties of items. IRT models yield item and latent

trait estimates within a non-linear transformation of the raw score that does not vary with the characteristics of the population.

When working with IRT models, it is important to determine that the latent construct measured by all the items is statistically unidimensional, which means all items measure only one latent trait (Hambleton et al., 1991). Research shows that ADAS-cog demonstrates strong unidimensionality through different factor analytic techniques (Benge et al., 2009). ADAS-cog also satisfies other assumptions of IRT such as invariance and local independence (Verma et al., 2015). Therefore, IRT could be applied to the analysis of ADAS-cog to approximate the probability of each category of items with the person parameters and item parameters estimated from the sample.

Given the probability of responses estimated by person and item parameters in the IRT model, we can calculate the mutual information between each item and the latent trait in a scale. Items with high mutual information are supposed to be more sensitive to the disease progression.

This work investigated the IRT-based mutual information of each item in ADAS-cog to select sensitive items in MCI population. To examine if the mutual information well measured the sensitivity to disease severity change, we estimated its correlation with the rate of change within two years of each items on the same population. Other IRT-based statistics, such as the Fisher information, were compared. A composite score based on the item of high IRT-based mutual information was also compared with other subscales of ADAS-cog.

Materials and method

Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and

early Alzheimer's disease (AD) (Weiner et al., 2012). All data points available for MCI subjects enrolled in ADNI-1, ADNI Go and ADNI-2 were included in this analysis.

Eight hundred and sixty-six MCI subjects were included in this study. Among them, 308 subjects were diagnosed as Early MCI (EMCI), 558 of them were diagnosed as Late MCI (LMCI) at baseline; 512 subjects were males, and 354 subjects were female. Their ages were ranged from 54 to 91. All subjects had APOE information collected at baseline: 430 subjects do not have APOE4 allele. 338 subjects have one APOE4 allele, and 94 subjects have two alleles.

In this study, we evaluated the ADAS-cog measurement scale and its subscales in ADNI. The 13 items in ADAS-cogs are Word Recall (Q1), Commands (Q2), Construction (Q3), Delayed Word Recall (Q4), Naming (Q5), Ideational Praxis (Q6), Orientation (Q7), Word Recognition (Q8), Recall Instruction (Q9), Spoken Language (Q10), Word Finding Difficulty (Q11), Comprehension (Q12), and Number Cancellation (Q14).

Item Response Theory

Item Response Theory (IRT) (Woods and Baker, 1985) comprised of mathematical models describing the association between a respondent's underlying level on a continuum of the latent trait and the probability of a particular item response using a nonlinear function. The different IRT models are distinguished by the functional form specified for the relationship between latent trait and item response probability. Items could be dichotomous or polytomous (item with multiple response categories). There are three main types of dichotomous IRT models. The **One parameter model** (1-PL) or Rasch Model allows items to vary in their difficulty level (β) but equally discriminated. The **Two parameter model** (2-PL) extends the 1-PL Rasch model by estimating an item discrimination parameters (α) qualifying how well the item distinguishes subjects with different latent levels. **Three parameter model** (3-PL) includes a pseudo-guessing parameter (c). In an ability testing, subjects can get an answer by chance. This parameter adjusts for the impact of chance on observed scores.

Since the items in the ADAS-cog are widely varied in terms of categories number and setting, they are assumed to be not equally discriminating. The responses of psychometric tests in ADNI

were mostly provided by examiners. Thus the impact of chance on the response is not considered. For these reasons, the 2-PL model could be adapted to describe the ADAS-cog data.

The **Graded Response Model** (GRM) is an extension of the 2-PL model which is appropriate to use on polytomous items (Samejima, 1970). In the GRM, items need not have the same number of response categories. Each item is described by a discrimination parameter α and between category threshold parameters β, which represent the trait level necessary to respond above threshold with 0.5 probabilities. In GRM, for item k, the probability of the person i to endorse the lth or higher response categories is given by

$$P(X_{ik} \geq l|\theta_i) = \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{kl})}},$$

where the X_{ik} is the ordinal manifest variable with L_k possible response categories. The threshold β_{kl} can be considered as the difficulty of responding with category l or higher for item k with $\beta_{k1} < \cdots < \beta_{kl} < \beta_{k,L_k-1}$ and $\beta_{k,L_k} = \infty$. θ_i is the person parameter of the person i.

Therefore, the probability of observing lth categories is

$$P(X_{ik} = l|\theta_i) = P(X_{ik} \geq l|\theta_i) - P(X_{ik} \geq l+1|\theta_i)$$

$$= \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{kl})}} - \frac{1}{1 + e^{-\alpha_k(\theta_i - \beta_{k,l+1})}}$$

where we take $P(X_{ik} \geq 0|\theta_i) = 1$and α_k is the discrimination parameter of the item, β_{kl} is the category threshold parameter.

Mutual information

In information theory, one of the key concepts is entropy, a measure of disorder or uncertainty (Commenges, 2015; Cover and Thomas, 2006). It is found to be the only appropriate function to measure the information for the observation of a random variable X taking different values x and having a distribution f.

The entropy is defined as:

$$H(X) = \sum_{x \in X} f(x) \log \frac{1}{f(x)} = -\sum_{x \in X} f(x)\ \log f(x)$$

For two dependent variables, the information needed to describe the outcome of a random variable X given the value of another random variable Y is qualified by the conditional entropy

$$H(X|Y) = -\sum_{y \in Y} f(y) \sum_{x \in X} f(x|y)\ \log f(x|y)$$

Then, the amount of information contained in one random variable X about the other random variable Y can be qualified by the mutual information

$$I(X;Y) = H(X) - H(X|Y)$$

Note that the mutual information is symmetric. If X and Y are independent, the mutual information is null. Therefore $I(X;Y)$ can be considered as a measure of dependence.

In the context of this study, the mutual information contained in item k about the latent trait θ is of concerned:

$$I(X_k;\theta) = H(X_k) - H(X_k|\theta)$$

where

$$H(X_k) = -\sum_{l=1}^{L_k} P(X_k = l) \log P(X_k = l)$$

given

$$P(X_k = l) = \int_{-\infty}^{\infty} P(X_k = l|\theta)\, \varphi(\theta) d\theta$$

assuming $\theta \sim N(0,1)$

and

$$H(X_k|\theta) = -\int_{-\infty}^{\infty} \sum_{l=1}^{L_k} P(X_k = l|\theta) \log P(X_k = l|\theta)\ \varphi(\theta) d\theta$$

Statistical analysis

Mutual information estimation and comparison

The amount of IRT-based mutual information of the items in ADAS-cog was estimated using the baseline data of MCI patients in ADNI 1, 2 and Go.

Then the mutual information of each item was compared with other IRT-based statistics. The application of IRT provides other statistics representing the characters of the item, such as the discrimination and Fisher information. They could also be utilized in the item selection. Discrimination of items (parameter α_k in the GRM model for item k) illustrates the capacity of an item to discriminate between contiguous trait levels near the inflection point. More discriminating items provide greater information about the latent trait than less discriminating items. In the research of IRT, the Fisher information is the most used type of information. IRT-based Fisher information equals to the variance explained, showing how effectively a measure captures the latent trait. It also provides the precision of measure. For an item k with response X_k, Fisher information is defined as (Lord, 1980)

$$I_k^{Fisher} = -E\left(\frac{\partial^2 \operatorname{Log} P(X_k = l|\theta)}{\partial \theta^2}\right)$$

where $P(X_k = l|\theta)$ is the conditional probability of X_i given θ estimated by the GRM. These two statistics have been proposed for evaluation of the item

Since the mutual information shows the dependence of the item to the latent trait, it is supposed that an item with high mutual information is more sensitive to the disease severity change. The discrimination and Fisher information for each item in ADAS-cog were estimated based on the baseline data of MCI subjects in ADNI. Then they were compared with mutual information by

their correlation with the rate of change of these items in follow-up data to see which statistics correspond better to the disease severity change in time.

The rates of change were estimated using the two years of follow-up data of the same MCI patients. The longitudinal linear mixed-effects model was fitted to Z-score transformed of each item with the fixed effect of time and the random effect of subjects. This model used the following covariates: age, gender, education level and APOE4 status. Missing data were assumed to be missing at random, and no values were imputed for missing data. The ADAS-cog scores at baseline, 12 months and 24 months of the MCI patients were included in the estimation of the rate of change.

Composite score

To reduce the variability and enhance the reliability, it is of interest to select the more sensitive sub-items in the scale. The composite score of the items with high mutual information could be a sensitive measurement of disease severity change. To keep the facility of utilization, the scores of these high mutual information items were simply summed up to generate the composite score. This composite score was compared with other composite scores of ADAS-cog including the ADAS-cog11, ADAS-cog13 and also the composite score of the items Word Recall (Q1), Delayed Word Recall (Q4) and Orientation (Q7), which were selected by Huang *et al.* (Huang et al., 2014) and Raghavan *et al*. (Wouters et al., 2008) as sensitive items for MCI subjects evaluation. Their rates of change were estimated using the linear mixed-effected model on the two years follow-up data of the MCI patients in ADNI for the comparison.

Implementation

We used the *R* statistical computing platform version 3.02. The summary data of ADNI were obtained by the package *adnimerge*. The application of Graded Response Model, the estimation of discrimination and FI of each item was conducted by *R* package *ltm*. The mutual information was estimated with custom-made *R* functions. The linear mixed–effects model was applied with *R* package *lme4.*

Results

Mutual information and comparison

The mutual information of the 13 items in ADAS-cog was estimated using baseline data from all MCI subjects in ADNI 1, ADNI Go and ADNI 2. The processing time for this calculation was 6.82 seconds with R 3.02, 64-bit operation system with 16G RAM.

The ranking of items by the mutual information amount that they contain could be found in **Table 1**. The items Delayed Word Recall, Word Recall and Word Recognition exhibited obviously higher IRT-based mutual information compared to other items. Orientation has slightly higher mutual information in the rest of items. The mutual information of the rest of items in ADAS-cog was very close (between 0.02 to 0.06). The ranking the items based on the Fisher information and Discrimination could also be found in **Table 1**.

Mutual Information		**Fisher Information**		**Discrimination**	
Item	Value	Item	Value	Item	Value
Delayed Word Recall (Q4)	0.78	Word Recall (Q1)	16.97	Delayed Word Recall (Q4)	3.01
Word Recall (Q1)	0.58	Delayed Word Recall (Q4)	15.18	Word Recall (Q1)	2.65
Word Recognition (Q8)	0.29	Word Recognition (Q8)	4.68	Word Recognition (Q8)	1.26
Orientation (Q7)	0.08	Orientation (Q7)	1.86	Recall Instruction (Q9)	0.79
Naming (Q5)	0.06	Ideational Praxis (Q6)	1.5	Ideational Praxis (Q6)	0.77
Number Cancellation (Q14)	0.06	Recall Instruction (Q9)	1.42	Orientation (Q7)	0.73
Ideational Praxis (Q6)	0.05	Naming (Q5)	1.34	Naming (Q5)	0.68
Comprehension (Q12)	0.03	Number Cancellation (Q14)	1.19	Comprehension (Q12)	0.61
Construction (Q3)	0.03	Comprehension (Q12)	0.84	Number Cancellation (Q14)	0.55
Recall Instruction (Q9)	0.03	Construction (Q3)	0.58	Spoken Language (Q10)	0.42
Commands (Q2)	0.02	Spoken Language (Q10)	0.4	Construction (Q3)	0.4
Spoken Language (Q10)	0.02	Commands (Q2)	0.39	Commands (Q2)	0.39
Word Finding Difficulty (Q11)	0.02	Word Finding Difficulty (Q11)	0.31	Word Finding Difficulty (Q11)	0.31

Table 1. The values of mutual information, Fisher Information and Discrimination of the 13 items in ADAS-cog and their ranking accordingly.

The estimated Fisher information presented a larger range of value (0.31 to 16.97) compared to Mutual Information (0.03 to 0.78) and Discrimination (0.31 to 3.01). Nevertheless, there exists similarity in the three statistics: items Delayed Word Recall, Word Recognition and Word Recall were the top three highly ranked. Word Finding Difficulty was the last for three different ranking.

The Mutual Information, Fisher Information and Discrimination were compared with the rate of change (Z-score) of 13 items in ADAS-cog estimated by the longitudinal linear mixed-effects model using the follow-ups data (12 months, 24 months) of the baseline MCI subjects in ADNI. The correlations between the rate of change and different statistics of items were shown on the heat map of Pearson's correlation coefficients (**Figure 1**). Although the three statistics estimated using baseline data were similar (their correlations were between 0.0.97- 0.98), the IRT-based mutual information of the items were more correlated with the rate of change estimated on the two years of follow-up data ($\rho = 0.88$) compared to that of the Fisher Information ($\rho = 0.78$) and the discrimination ($\rho = 0.83$).

Figure 1. The Pearson's correlation coefficient between the mutual information, Fisher Information and Discrimination estimated on the baseline MCI patient sample and the rate of change estimated on the two years of follow-up data.

Composite score

The three items with the highest mutual information (Word Recall, Word Recognition and Delayed Word Recall) were used to develop a composite score (High mutual information). They were compared with the ADAS-cog11, ADAS-cog13 and also the composite score of the items Word Recall, Delayed Word Recall and Orientation (Composite).

The rate of change of these five composite scores estimated by the linear mixed–effects model on two years of follow-ups data were illustrated in **Figure 2**. With the complete set of items, ADAS-cog13 exhibits the highest rate of change in two years (0.30). The composite score of high mutual information items shows a higher rate of change (0.26) compared to the rate of change of ADAS-cog11 (0.18) and the composite score of the sensitive items selected in other studies (0.16)

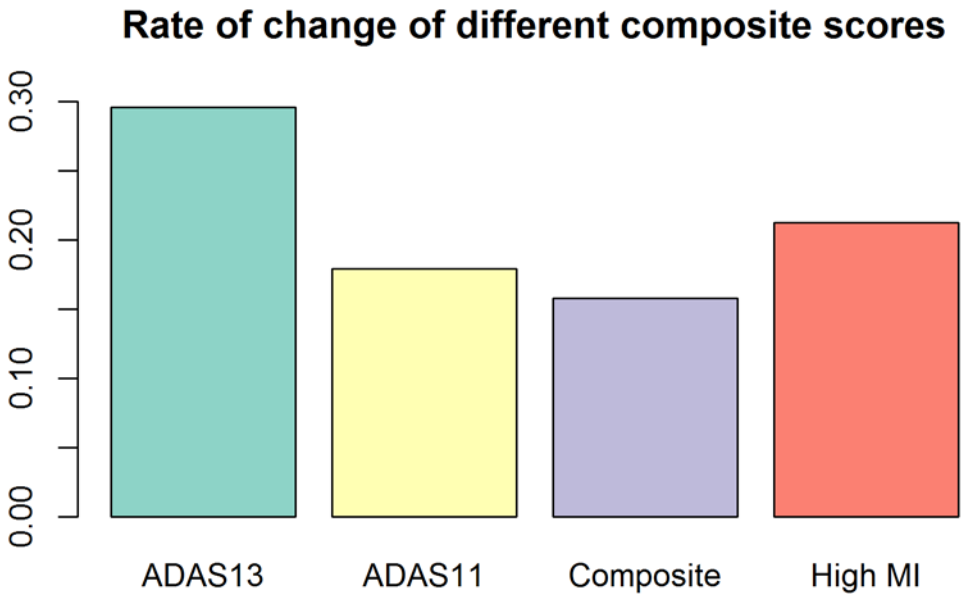

Figure 2 The rate of change (Z-score) through two years follow-up data of the subscales of the ADAS-cog. **ADAS 11**: The subscale of ADAS-cog 13 without item Delayed Word Recall and Number Cancellation.

Composite: items Word Recall, Delayed Word Recall and Orientation. **High mutual information**: Word Recall, Word Recognition and Delayed Word Recall.

Discussion

This study is the first attempt to apply the mutual information on the item selection for the assessment of cognitive. Mutual information has been widely used in the telecommunication and the machine learning as a feature selection method, whereas the exploration of its application on measure scale, especially in combination with the IRT theory, is yet to start.

In this article, we proposed to use the IRT-based mutual information to select sensitive items in cognitive measurement such as ADAS-cog. ADNI 1, ADNI Go and ADNI 2 baseline and two years of follow-ups data were used to identify items which are sensitive to disease severity change in the MCI population. The items Word Recall (Q4), Delayed Word Recall (Q1) and Word Recognition (Q8) showed high mutual information in the result. This list is highly overlapped to those identified in previously reported analyses. The items Word Recall, Delayed Word Recall were selected as most sensitive items in ADAS-cog by Huang *et al*. (Huang et al., 2014) using the signal-to-noise ratio for detecting the hypothetical treatment effect based on the longitudinal data and also by Ueckert *et al*. using the IRT-based Fisher information (Ueckert et al., 2014). Hannesdottir *et al.* proposed a subset of the ADAS-cog for prodromal AD patients as being the most sensitive based on ADNI including these two items (Hannesdottir and Snaedal, 2002). They were also found to have the largest amount of change across clinical categories in the research of Raghavan *et al.* (Raghavan et al., 2013). In a composited score proposed by Wang *et al.*, Delayed Word Recall and Word Recognition were selected by a linear regression model on MCI subjects (Wang et al., 2016). To sum up, each item having high mutual information has been identified separately in several studies.

The items with high mutual information are related to the memory impairment evaluation. The ADAS-cog is originally designed to assess three different cognitive domains including memory, language and praxis (Rosen et al., 1984). The top four items with high mutual information are exactly the four items measuring memory impairment: Word Recall (Q4), Delayed Word Recall (Q1) and Word Recognition (Q8) and Orientation (Q7). Studies suggest that impairment in the

memory, language and praxis cognitive domains progress differently based on the brain regions involved in different stages of Alzheimer's disease (Frisoni et al., 2008; Thompson et al., 2003). Memory loss starts earlier in the pathology of Alzheimer's disease and has been considered characteristic. The items related to the memory also showed higher sensibility in the drug effect evaluation (Verma et al., 2015). These studies may also indicate that the high mutual information items could be sensitive markers for the early stage of the disease such as MCI.

Furthermore, previous analysis identified that items Commands, Construction, Naming, Praxis, Recall Instruction, Language, Word Finding Difficulty and Comprehension exhibit ceiling effects in virtually every cohort on MCI subjects (Raghavan et al., 2013). Nevertheless, none of these items are selected by this method proposed. This may show that the application of mutual information helps to eliminate uninformative items in a scale by only using baseline data.

Not only coherent with previous results, but the items with high IRT-based mutual information also showed more sensitivity to change through different comparisons. For an individual item, when compared to other IRT-based statistics (Fisher information and item discrimination), their mutual information was better correlated to the rate of change of the two years of follow-up data. In the comparison of the subscales of ADAS-cog, the rate of change in two years of the composite scores which summed only the 3 items with high mutual information were larger than that of the composite score of items Q1, Q4, and Q7 selected by Wang *et al.* (Huang et al., 2014) *and* Ueckert *et al*. (Ueckert et al., 2014) and the ADAS-cog 11. This result reinforced our viewpoint that the IRT-based mutual information could be a good statistic for sensitivity evaluation. The large reduction of items number would also benefit the reduction of variability of the scale.

This study has still some limitations that should be mentioned. Firstly, although ADNI is a quality study with large sample size, there still exists lacking in the data. In the 866 MCI subjects who have ADAS-cog score in the baseline, 785 of them have a visiting point at 12^{th} month, 674 of them have a visiting point at 24^{th} month. The reduction of sample size compared to the baseline would decrease the accuracy of the rate of change estimation. Merely 561 subjects have a visiting point at 36^{th} month. Therefore only the two years follow-up data were included in this study. Secondly, we focused on ADNI data to identify sensitive items. However, validation of the result using another dataset is also important. Reliability of the selection of sensitive items should be

assessed when more data are available. Thirdly, the items with high mutual information are slightly different compared to the previous two studies of sensitive items selection using ADNI data (Huang et al., 2014; Ueckert et al., 2014): the inclusion of item Word Recognition instead of Orientation. The divergence of the results could be led by the different data source used in these analyses. For example, the ADNI 2 data is not included in these two studies. After all, Orientation is the item with the fourth highest mutual information in ADAS-cog.

The selection of sensitive items to disease severity change in cognitive measurement remains an active topic. In this study, we proposed the IRT-based mutual information criteria to evaluate the sensitivity of items. The calculation is fast (0.52 second per item on 866 samples) and only baseline data are sufficient. This property could be of interest in the clinical researches in which the trial design may be adapted to the result of the baseline data. It correlates well to the long-term rate of change compared to other statistics and the selected items correspond to several previous studies. Furthermore, this method could be extended to the item selection on other cognitive assessment such as the MMSE. We believe that this method could help to improve power for MCI trials and eventually endpoint development in different diseases research.

References

Ard, M.C., Galasko, D.R., and Edland, S.D. (2013). Improved Statistical Power of Alzheimer Clinical Trials by Item-Response Theory. Alzheimer Dis. Assoc. Disord. , 187–191.

Balsis, S., Unger, A.A., Benge, J.F., Geraci, L., Doody, R.S.R.S., Rosen, W.G., Mohs, R.C., Davis, K.L., Rockwood, K., Fay, S., et al. (2012). Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. Alzheimers. Dement. , 288–294.

Benge, J.F., Balsis, S., Geraci, L., Massman, P.J., and Doody, R.S. (2009). How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease? Dement. Geriatr. Cogn. Disord. , 63–69.

Cano, S.J., Posner, H.B., Moline, M.L., Hurt, S.W., Swartz, J., Hsu, T., and Hobart, J.C. (2010). The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. J. Neurol. Neurosurg. Psychiatry , 1363–1368.

Commenges, D. (2015). Information Theory and Statistics: an overview.

Cooper, M.L. (2000). Information measures for object recognition accommodating signature variability. IEEE Trans. Inf. Theory , 1896–1907.

Cover, T.M., and Thomas, J.A. (2006). Elements of information theory (Wiley-Interscience).

Frisoni, G.B., Henneman, W.J., Weiner, M.W., Scheltens, P., Vellas, B., Reynish, E., Hudecova, J., Hampel, H., Burger, K., Blennow, K., et al. (2008). The pilot European Alzheimer's Disease Neuroimaging Initiative of the European Alzheimer's Disease Consortium. Alzheimers Dement , 255–264.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). Fundamentals of item response theory. Contemp. Sociol. , 289.

Hannesdottir, K., and Snaedal, J. (2002). A study of the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) in an Icelandic elderly population. Nord. J. Psychiatry , 201–206.

Huang, Y., Ito, K., Billing, C.B., and Anziano, R.J. (2014). Development of a straightforward and

sensitive scale for MCI and early AD clinical trials. Alzheimer's Dement. , 404–414.

Ihl, R., Frölich, L., Dierks, T., Martin, E.-M., and Maurer, K. (1992). Differential validity of psychometric tests in dementia of the Alzheimer type. Psychiatry Res. , 93–106.

Ihl, R., Grass-Kapanke, B., Jänner, M., and Weyer, G. (1999). Neuropsychometric Tests in Cross Sectional and Longitudinal Studies - A Regression Analysis of ADAS - Cog, SKT and MMSE. Pharmacopsychiatry , 248–254.

Karasuyama, M., and Sugiyama, M. (2012). Canonical dependency analysis based on squared-loss mutual information. Neural Networks , 46–55.

Koppel, J. (2005). Memory impairment, but not cerebrovascular disease, predicts progression of MCI to dementia. Neurology , 766; author reply 766.

Liu, C.-L. (2005). Using Mutual Information for Adaptive Item Comparison and Student Assessment. Educ. Technol. Soc. , 100–119.

Lord, F.M. (1980). Applications of item response theory to practical testing problems (L. Erlbaum Associates).

Raghavan, N., Samtani, M.N., Farnum, M., Yang, E., Novak, G., Grundman, M., Narayan, V., DiBernardo, A., and Alzheimer's Disease Neuroimaging Initiative (2013). The ADAS-Cog revisited: novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. Alzheimers. Dement. , S21-31.

Rosen, W.G., Mohs, R.C., and Davis, K.L. (1984). A new rating scale for Alzheimer's disease.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. Psychometrika , 139.

Silva, J., and Narayanan, S.S. (2009). Discriminative wavelet packet filter bank selection for pattern recognition. IEEE Trans. Signal Process. , 1796–1810.

Skinner, J., Carvalho, J.O., Potter, G.G., Thames, A., Zelinski, E., Crane, P.K., Gibbons, L.E., and Alzheimer's Disease Neuroimaging Initiative (2012). The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness

in MCI. Brain Imaging Behav. , 489–501.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics , S231–S240.

Thévenaz, P., Unser, M., and Thevenaz, P. (2000). Optimization of mutual information for multiresolution image registration. IEEE Trans. Image Process. , 2083–2099.

Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., et al. (2003). Dynamics of gray matter loss in Alzheimer's disease. J. Neurosci. , 994–1005.

Ueckert, S., Plan, E.L., Ito, K., Karlsson, M.O., Corrigan, B., Hooker, A.C., and Alzheimer's Disease Neuroimaging Initiative (2014). Improved Utilization of ADAS-Cog Assessment Data Through Item Response Theory Based Pharmacometric Modeling. Pharm. Res. , 2152–2165.

Verma, N., Beretvas, S.N., Pascual, B., Masdeu, J.C., Markey, M.K., Rosen, W., Mohs, R., Davis, K., Cano, S., Posner, H., et al. (2015). New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in clinical trials. Alzheimers. Res. Ther. , 64.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. Educ. Psychol. Meas. , 1017–1035.

Wang, J., Logovinsky, V., Hendrix, S.B., Stanworth, S.H., Perdomo, C., Xu, L., Dhadda, S., Do, I., Rabe, M., Luthman, J., et al. (2016). ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J. Neurol. Neurosurg. Psychiatry jnnp-2015-312383.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al. (2012). The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's Dement. , S1--S68.

Westover, M.B., and O'Sullivan, J.A. (2008). Achievable rates for pattern recognition. IEEE Trans. Inf. Theory , 299–320.

Weyer, G., Erzigkeit, H., Kanowski, S., Ihl, R., and Hadler, D. (1997). Alzheimer's Disease

Assessment Scale: reliability and validity in a multicenter clinical trial. Int. Psychogeriatrics , 123–138.

Woods, A., and Baker, R. (1985). Item response theory. Lang. Test. , 117–140.

Wouters, H., van Gool, W.A., Schmand, B., and Lindeboom, R. (2008). Revising the ADAS-cog for a more accurate assessment of cognitive impairment. Alzheimer Dis. Assoc. Disord. , 236–244.

Chapter 5

Discussion

As an alternative to the classical test theory using the directly summed scores of a test, Item Response Theory that fully considers the difference of the item properties has been widely applied to categorical data analysis. Aiming to deepen our understanding to the disease and promote the disease-modifying therapies, categorical data such as GWAS data and rating scales scores are widely utilized in neurodegenerative diseases researches. In this thesis, after the studies of the different aspects of the IRT and its extensions, we applied the IRT models on the analysis of the categorical data. By a series of tests under the framework of Rasch analysis, we showed that except some flaws of the setting of certain items, the CMTNS scale is validated for CMT1A disease assessment, but more suitable for moderate to severe forms of the disease. To enhance its precision on mild-to-moderate severity assessment, more items and/or categories are needed. This study may help the endpoint selection on the clinical trials of CMT disease. Then Rasch model was applied on the GWAS data as a multi-marker genetic association test. This novel method has shown better performance compared to other association tests through simulations. Part of the genes found associated with the Alzheimer's disease by the proposed methods has not been mentioned in other GWAS studies of AD but functionally linked to the disease pathology. The results of pathway and network analysis of these genes also show correspondence to the known AD pathology. The Rasch model based genetic association test may gain an insight into the mechanism of AD. Finally, the Item Response Theory was combined with the mutual information to evaluate the sensitivity of the items in the ADAS-cog scale. Compared to other IRT-based statistics, this IRT-based mutual information of items was better correlated with their rate of change. The composite score of the items with high mutual information showed a higher rate of change compared to that of other composite scores of ADAS-cog subscales. This result may help

to reduce the variability for predementia clinical researches and increase the test procedure efficiency.

The research work that I presented in this thesis also suggests some interesting research perspectives.

Brain imaging data analysis

The loss of brain neurons is a primary symptom of central nervous system neurodegenerative diseases, such as Alzheimer's disease, Parkinson's disease and Huntington's disease. Numerous studies suggested that the change in brain structure (detected by MRI) or brain glucose metabolism (detected by FDT-PET) had higher statistical power to detect progression of disease than clinical or cognitive measures (Weiner et al., 2010). Firstly, neuroimaging has a higher reliability than cognitive measures and thus greatly increases the power to detect longitudinal change and treatment effects. Secondly, neuroimaging has "face validity" as an index of disease progression, because generally it is accepted that loss of synapses and neurons is a result of neurodegeneration. Finally, FDG-PET and MRI imaging are validated quantitatively to some extent by correlation with cognitive and functional measure and neuropathology at autopsy (Mueller et al., 2005b). Besides GWAS data and clinical rating scales data, the brain imaging data can complement the information gained from clinical measures and thus is worth to be concerned. In a side project of collaboration with Fudan University (Shanghai, China), we have validated the association between a SNP on a solute carrier transporter gene and putamen volume on the sample from the Three City Project. The brain imaging data could be integrated into the application of the methods developed in this thesis. For instance, in the study of Rasch model based multi-marker genetic association test, the volume of certain Region Of Interest (RIO), such as hippocampus, can be used as a quantitative trait in the test instead of the binary trait of case and control. The association of the multi-marker Rasch genetic score to the disease can then simply assessed with a linear model. In this case, novel genes related to the brain atrophy rate of the Alzheimer's disease may be discovered.

Histogram based Mutual information estimation

In the study of ADAS-cog item selection previously presented in Chapter 4, we used the mutual information as criteria of the item sensitivity. The GRM model in the IRT was used to approximate the probability of density functions of each category in an item. The mutual information (MI) of an item is then the sum of the IRT-based mutual information for each category assuming a normal distribution of the person ability parameters. The IRT model provides a precise estimation of the category endorsement and the IRT-based MI is higher correlated to the disease progression. However, it is not the only way to combine the IRT and information theory. There are several approaches to estimating the MI from finite samples. One of the simplest methods is the histogram-based method (Moddemeijer, 1989), which partitions the space into several bins and count the number of elements in each bin. This method is very easy and efficient from the computational point of view. With this method, we can first estimate the ability parameter for each individual θ using IRT model instead of assuming that they are normal distributed. The space of the person ability parameters θ can be divided into multiple bins. The $P(X_{ik} = \mathrm{x}|\theta_i)$ for each bin i can be approximated using the $\hat{P}(X_{ik} = \mathrm{x}|\theta_i)$ according to the category selection of the individuals in that range of θ. The $P(X_{ik} = \mathrm{x})$ can be estimated by the margin probability of $P(X_{ik} = \mathrm{x}|\theta_i)$. Then we can calculate the MI based on this histogram-based probability. The result can be compared with the IRT based MI of ADAS-cog items and may reveal other items sensitive to disease progression.

Other applications

The methods proposed in this thesis can have further applications to the data analysis of neurodegenerative diseases and other diseases.

In the study of Rasch analysis on the CMTNS, we developed a set of functions with *R* language to evaluate the psychometrical properties of the scale and its items comprehensively. A new *R* package of Rasch analysis can be therefore constructed afterward based on these functions. To facilitate the analysis for non-statisticians, an automate pipeline may also be built in this package to reproduce a Rasch analysis report giving the raw data of a rating scale.

The Rasch Model based association test also opens other opportunities for GWAS data analysis. First, the analysis of multiple markers is not only on the gene level, but can also be applied to the analysis of the SNPs on the whole pathway, knowing that disease susceptibility is actually likely

to depend on the cumulative effect of multiple variants in several genes interacting in functional pathways (Lehne et al., 2011). The pathway analysis also allows the consideration of enriched biological information. Second, besides in the association test, the genetic score estimated by Rasch model could also be used as a predictor of the disease for the classification of patients using machine learning methods. Third, the disease risk may be determined by multiple rare mutations (Madsen and Browning, 2009). The Rasch model is suitable for the inclusion of rare variants, as most rare variants analyses focus on gene level test by collapsing the effects of all rare SNPs in a gene into a single test of association (Fridley and Biernacka, 2011).

The IRT-based MI that we developed can be expanded to the item sensitivity analysis of other clinical scales, such as the MMSE and CDR, which are important in the disease evaluation. It could be eventually be applied to the whole neuropsychological battery in the study of a certain disease to select the most sensitive items across different tests to develop a composite score.

In conclusion, understanding the pathological mechanisms of neurodegenerative disease and developing new treatment remain challenging. For these purposes, the adaptation of IRT on the diseases data is obviously necessary to provide a thorough analysis. Through this thesis, we discovered several usages of the IRT. It provides us the opportunities to identify the defects of the CMTNS scale, discover more AD-associated genes, and select sensitive items in ADAS-cog. It has great potential to improve the data analysis in the health care. Still, further statistical developments are needed to be able to fully exploit and analyze the categorical data available in the disease studies.

References

Andersen, E.B. (1973). A goodness of fit test for the rasch model. Psychometrika *38*, 123–140.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika *43*, 561–573.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? Med. Care *42*, I7-16.

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the Polytomous Rasch Model. Psychometrika *75*, 292–308.

Ard, M.C., Galasko, D.R., and Edland, S.D. (2013). Improved Statistical Power of Alzheimer Clinical Trials by Item-Response Theory. Alzheimer Dis. Assoc. Disord. *27*, 187–191.

Attarian, S., Vallat, J.-M., Magy, L., Funalot, B., Gonnaud, P.-M., Lacour, A., Péréon, Y., Dubourg, O., Pouget, J., Micallef, J., et al. (2014). An exploratory randomised double-blind and placebo-controlled phase 2 study of a combination of baclofen, naltrexone and sorbitol (PXT3003) in patients with Charcot-Marie-Tooth disease type 1A. Orphanet J. Rare Dis. *9*, 199.

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., Jones, E., WHO, Ferri, C., Prince, M., Brayne, et al. (2011). Alzheimer's disease. Lancet (London, England) *377*, 1019–1031.

Balsis, S., Unger, A.A., Benge, J.F., Geraci, L., Doody, R.S.R.S., Rosen, W.G., Mohs, R.C., Davis, K.L., Rockwood, K., Fay, S., et al. (2012). Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. Alzheimers. Dement. *8*, 288–294.

Bateman, R.J., Xiong, C., Benzinger, T.L.S., Fagan, A.M., Goate, A., Fox, N.C., Marcus, D.S., Cairns, N.J., Xie, X., Blazey, T.M., et al. (2012). Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. N. Engl. J. Med. *367*, 795–804.

Benge, J.F., Balsis, S., Geraci, L., Massman, P.J., and Doody, R.S. (2009). How well do the ADAS-cog and its subscales measure cognitive dysfunction in Alzheimer's disease? Dement.

Geriatr. Cogn. Disord. *28*, 63–69.

Birks, J., and Harvey, R.J. (2006). Donepezil for dementia due to Alzheimer's disease. In Cochrane Database of Systematic Reviews, J. Birks, ed. (Chichester, UK: John Wiley & Sons, Ltd), p. CD001190.

Bond, T.G., and Fox, C.M. (2013). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, Second Edition.

Bouaziz, M. (2012). PhD Thesis Statistical methods to account for di erent sources of bias in Genome-Wide association studies.

Bredesen, D.E., Rao, R. V., and Mehlen, P. (2006). Cell death in the nervous system. Nature *443*, 796–802.

Brown, R.C., Lockwood, A.H., and Sonawane, B.R. (2005). Neurodegenerative diseases: An overview of environmental risk factors.

Burns, A., and Iliffe, S. (2009). Alzheimer's disease. BMJ *338*, b158.

Burns, J., Ouvrier, R., Estilow, T., Shy, R., Laurá, M., Pallant, J.F., Lek, M., Muntoni, F., Reilly, M.M., Pareyson, D., et al. (2012). Validation of the Charcot-Marie-Tooth disease pediatric scale as an outcome measure of disability. Ann. Neurol. *71*, 642–652.

Cano, S.J., Posner, H.B., Moline, M.L., Hurt, S.W., Swartz, J., Hsu, T., and Hobart, J.C. (2010). The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. J. Neurol. Neurosurg. Psychiatry *81*, 1363–1368.

Christensen, K.B., Bjorner, J.B., Kreiner, S., and Petersen, J.H. (2002). Testing unidimensionality in polytomous Rasch models. Psychometrika *67*, 563–574.

Commenges, D. (2015). Information Theory and Statistics: an overview.

Cooper, M.L. (2000). Information measures for object recognition accommodating signature variability. IEEE Trans. Inf. Theory *46*, 1896–1907.

Cover, T.M., and Thomas, J.A. (2006). Elements of information theory (Wiley-Interscience).

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*, 297–334.

Dyck, P.J., and Lambert, E.H. (1968). Lower motor and primary sensory neuron diseases with peroneal muscular atrophy. II. Neurologic, genetic, and electrophysiologic findings in various neuronal degenerations. Arch. Neurol. *18*, 619–625.

Ekins, S., Litterman, N.K., Arnold, R.J.G., Burgess, R.W., Freundlich, J.S., Gray, S.J., Higgins, J.J., Langley, B., Willis, D.E., Notterpek, L., et al. (2015). A brief review of recent Charcot-Marie-Tooth research and priorities. F1000Research *4*, 53.

Förstl, H., and Kurz, A. (1999). Clinical features of Alzheimer's disease. Eur. Arch. Psychiatry Clin. Neurosci. *249*, 288–290.

French, G.A., and Dodd, B.G. (1999). Parameter recovery for the rating scale model using PARSCALE. J. Outcome Meas. *3*, 176–199.

Fridley, B.L., and Biernacka, J.M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. Eur. J. Hum. Genet. *19*, 837–843.

Frisoni, G.B., Henneman, W.J., Weiner, M.W., Scheltens, P., Vellas, B., Reynish, E., Hudecova, J., Hampel, H., Burger, K., Blennow, K., et al. (2008). The pilot European Alzheimer's Disease Neuroimaging Initiative of the European Alzheimer's Disease Consortium. Alzheimers Dement *4*, 255–264.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiske, A., Pedersen, N.L., HC, H., T, B., et al. (2006). Role of Genes and Environments for Explaining Alzheimer Disease. Arch. Gen. Psychiatry *63*, 168.

Gess, B., Röhr, D., Fledrich, R., Sereda, M.W., Kleffner, I., Humberg, A., Nowitzki, J., Strecker, J.-K., Halfter, H., and Young, P. (2011). Sodium-dependent vitamin C transporter 2 deficiency causes hypomyelination and extracellular matrix defects in the peripheral nervous system. J. Neurosci. *31*, 17180–17192.

Graham, R.C., and Hughes, R.A.C. (2006). A modified peripheral neuropathy scale: the Overall Neuropathy Limitations Scale. J. Neurol. Neurosurg. Psychiatry *77*, 973–976.

Green, K., and Frantom, C. (2002). Survey development and validation with the Rasch model. Int. Conf. Quest. Dev. Eval. Test. 1–30.

Grundman, M., Petersen, R.C., Ferris, S.H., Thomas, R.G., Aisen, P.S., Bennett, D.A., Foster, N.L., Jack, C.R., Galasko, D.R., Doody, R., et al. (2004). Mild Cognitive Impairment Can Be Distinguished From Alzheimer Disease and Normal Aging for Clinical Trials. Arch. Neurol. *61*, 59.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). Fundamentals of item response theory. Contemp. Sociol. *21*, 289.

Hannesdottir, K., and Snaedal, J. (2002). A study of the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) in an Icelandic elderly population. Nord. J. Psychiatry *56*, 201–206.

Hanson, B.A. (1998). Uniform DIF and DIF Defined by Differences in Item Response Functions. J. Educ. Behav. Stat. *23*, 244–253.

Harding, A.E., and Thomas, P.K. (1980a). Autosomal recessive forms of hereditary motor and sensory neuropathy. J. Neurol. Neurosurg. Psychiatry *43*, 669–678.

Harding, A.E., and Thomas, P.K. (1980b). The clinical features of hereditary motor and sensory neuropathy types I and II. Brain *103*, 259–280.

Hardy, J., and Allsop, D. (1991). Amyloid deposition as the central event in the aetiology of Alzheimer's disease. Trends Pharmacol. Sci. *12*, 383–388.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat.Genet. *41*, 1088–1093.

Hobart, J.C., Cano, S.J., Zajicek, J.P., and Thompson, A.J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol. *6*, 1094–1105.

Huang, Y., Ito, K., Billing, C.B., and Anziano, R.J. (2014). Development of a straightforward and sensitive scale for MCI and early AD clinical trials. Alzheimer's Dement. *11*, 404–414.

Ihl, R., Frölich, L., Dierks, T., Martin, E.-M., and Maurer, K. (1992). Differential validity of psychometric tests in dementia of the Alzheimer type. Psychiatry Res. *44*, 93–106.

Ihl, R., Grass-Kapanke, B., Jänner, M., and Weyer, G. (1999). Neuropsychometric Tests in Cross Sectional and Longitudinal Studies - A Regression Analysis of ADAS - Cog, SKT and MMSE. Pharmacopsychiatry *32*, 248–254.

Karasuyama, M., and Sugiyama, M. (2012). Canonical dependency analysis based on squared-loss mutual information. Neural Networks *34*, 46–55.

Kaya, F., Belin, S., Bourgeois, P., Micaleff, J., Blin, O., and Fontés, M. (2007). Ascorbic acid inhibits PMP22 expression by reducing cAMP levels. Neuromuscul. Disord. *17*, 248–253.

Kersten, P., White, P.J., and Tennant, A. (2014). Is the pain visual analogue scale linear and responsive to change? An exploration using rasch analysis. PLoS One *9*.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science (80-.). *308*, 385–389.

Koppel, J. (2005). Memory impairment, but not cerebrovascular disease, predicts progression of MCI to dementia. Neurology *64*, 766; author reply 766.

Kurtzke, J.F. (1983). Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology *33*, 1444–1452.

Lehne, B., Lewis, C.M., and Schlitt, T. (2011). From SNPs to genes: disease association at the gene level. PLoS One *6*, e20133.

DE LEON, M.J., MOSCONI, L., BLENNOW, K., DESANTI, S., ZINKOWSKI, R., MEHTA, P.D., PRATICO, D., TSUI, W., LOUIS, L.A.S., SOBANSKA, L., et al. (2007). Imaging and CSF Studies in the Preclinical Diagnosis of Alzheimer's Disease. Ann. N. Y. Acad. Sci. *1097*, 114–145.

Linacre, J. (2006). WINSTEPS Rasch measurement computer program. Chicago Winsteps. Com.

Liu, C.-L. (2005). Using Mutual Information for Adaptive Item Comparison and Student Assessment. Educ. Technol. Soc. *8*, 100–119.

Lord, F.M. (1980). Applications of item response theory to practical testing problems (L. Erlbaum Associates).

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell *66*, 219–232.

Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.

Mair, P., and Hatzinger, R. (2007). Extended Rasch Modeling : The R Package eRm.

Mandel, J., Bertrand, V., Lehert, P., Attarian, S., Magy, L., Micallef, J., Chumakov, I., Scart-Grès, C., Guedj, M., and Cohen, D. (2015). A meta-analysis of randomized double-blind clinical trials in CMT1A to assess the change from baseline in CMTNS and ONLS scales after one year of treatment. Orphanet J. Rare Dis. *10*, 74.

Mannil, M., Solari, A., Leha, A., Pelayo-Negro, A.L., Berciano, J., Schlotter-Weigel, B., Walter, M.C., Rautenstrauss, B., Schnizer, T.J., Schenone, A., et al. (2014). Selected items from the

Charcot-Marie-Tooth (CMT) Neuropathy Score and secondary clinical outcome measures serve as sensitive clinical markers of disease severity in CMT1A patients. Neuromuscul. Disord. *24*, 1003–1017.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika *47*, 149–174.

Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. Signal Processing *16*, 233–248.

Mudher, A., and Lovestone, S. (2002). Alzheimer's disease-do tauists and baptists finally shake hands? Trends Neurosci. *25*, 22–26.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C.R., Jagust, W., Trojanowski, J.Q., Toga, A.W., and Beckett, L. (2005a). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimers. Dement. *1*, 55–66.

Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., and Beckett, L. (2005b). The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. *15*, 869–877.

Murphy, S.M., Herrmann, D.N., McDermott, M.P., Scherer, S.S., Shy, M.E., Reilly, M.M., and Pareyson, D. (2011). Reliability of the CMT neuropathy score (second version) in Charcot-Marie-Tooth disease. J. Peripher. Nerv. Syst. *16*, 191–198.

Park, L.Q., Gross, A.L., McLaren, D.G., Pa, J., Johnson, J.K., Mitchell, M., Manly, J.J., and Alzheimer's Disease Neuroimaging Initiative, for the A.D.N. (2012). Confirmatory factor analysis of the ADNI Neuropsychological Battery. Brain Imaging Behav. *6*, 528–539.

Passage, E., Norreel, J.C., Noack-Fraissignes, P., Sanguedolce, V., Pizant, J., Thirion, X., Robaglia-Schlupp, A., Pellissier, J.F., and Fontés, M. (2004). Ascorbic acid treatment corrects the phenotype of a mouse model of Charcot-Marie-Tooth disease. Nat. Med. *10*, 396–401.

Patwardhan, M.B., McCrory, D.C., Matchar, D.B., Samsa, G.P., and Rutschmann, O.T. (2004). Alzheimer Disease: Operating Characteristics of PET— A Meta-Analysis. Radiology *231*, 73–80.

Potkin, S.G., Guffanti, G., Lakatos, A., Turner, J.A., Kruggel, F., Fallon, J.H., Saykin, A.J., Orro, A., Lupoli, S., Salvi, E., et al. (2009). Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease. PLoS One *4*, e6501.

Raghavan, N., Samtani, M.N., Farnum, M., Yang, E., Novak, G., Grundman, M., Narayan, V., DiBernardo, A., and Alzheimer's Disease Neuroimaging Initiative (2013). The ADAS-Cog revisited: novel composite scales based on ADAS-Cog to improve efficiency in MCI and early AD trials. Alzheimers. Dement. *9*, S21-31.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Reise, S.P., and Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. J. Educ. Meas. *27*, 133–144.

Revicki, D.A., Regulatory Issues and Patient-Reported Outcomes Task Force for the International Society for Quality of Life Research, Use, C. for M.P. for H., Kahnemann, D., Kruger, A., Schkade, D., Schwartz, N., Stone, A., Stone, A., Shiffman, S., et al. (2007). FDA draft guidance and health-outcomes research. Lancet (London, England) *369*, 540–542.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. J. Stat. Softw. *17*, 1–25.

Rogaeva, E. (2002). The solved and unsolved mysteries of the genetics of early-onset Alzheimer's disease. Neuromolecular Med *2*, 1–10.

Rosen, W.G., Mohs, R.C., and Davis, K.L. (1984). A new rating scale for Alzheimer's disease.

Rubinsztein, D.C. (2006). The roles of intracellular protein-degradation pathways in neurodegeneration. Nature *443*, 780–786.

Sadjadi, R., Reilly, M.M., Shy, M.E., Pareyson, D., Laura, M., Murphy, S., Feely, S.M.E., Grider, T., Bacon, C., Piscosquito, G., et al. (2014). Psychometrics evaluation of Charcot-Marie-Tooth Neuropathy Score (CMTNSv2) second version, using Rasch analysis. J. Peripher. Nerv. Syst. *19*, 192–196.

Samejima, F. (1970). Estimation of latent ability using a response pattern of graded scores. Psychometrika *35*, 139.

Schenone, A., Nobbio, L., Monti Bragadin, M., Ursino, G., and Grandis, M. (2011). Inherited Neuropathies. Curr. Treat. Options Neurol. *13*, 160–179.

Seshadri, S., Fitzpatrick, A.L., Ikram, M.A., DeStefano, A.L., Gudnason, V., Boada, M., Bis, J.C., Smith, A. V, Carassquillo, M.M., Lambert, J.C., et al. (2010). Genome-wide analysis of genetic loci associated with Alzheimer disease. JAMA *303*, 1832–1840.

Shaw, L.M., Korecka, M., Clark, C.M., Lee, V.M.-Y., and Trojanowski, J.Q. (2007). Biomarkers of neurodegeneration for diagnosis and monitoring therapeutics. Nat. Rev. Drug Discov. *6*, 295–303.

Silva, J., and Narayanan, S.S. (2009). Discriminative wavelet packet filter bank selection for pattern recognition. IEEE Trans. Signal Process. *57*, 1796–1810.

Skinner, J., Carvalho, J.O., Potter, G.G., Thames, A., Zelinski, E., Crane, P.K., Gibbons, L.E., and Alzheimer's Disease Neuroimaging Initiative (2012). The Alzheimer's Disease Assessment Scale-Cognitive-Plus (ADAS-Cog-Plus): an expansion of the ADAS-Cog to improve responsiveness in MCI. Brain Imaging Behav. *6*, 489–501.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J. Appl. Meas. *3*, 205–231.

Soto, C. (2003). Unfolding the role of protein misfolding in neurodegenerative diseases. Nat. Rev. Neurosci. *4*, 49–60.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics *18*, S231–S240.

Suárez-Falcón, J.C., and Glas, C.A.W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. Br. J. Math. Stat. Psychol. *56*, 127–143.

Tariot, P.N., Farlow, M.R., Grossberg, G.T., Graham, S.M., McDonald, S., Gergel, I., for the Memantine Study Group, for the M.S., DJ, S., JL, C., LE, H., et al. (2004). Memantine Treatment in Patients With Moderate to Severe Alzheimer Disease Already Receiving Donepezil. JAMA *291*, 317.

Tazir, M., Hamadouche, T., Nouioua, S., Mathis, S., and Vallat, J.-M. (2014). Hereditary motor and sensory neuropathies or Charcot–Marie–Tooth diseases: An update. J. Neurol. Sci. *347*, 14–22.

Thévenaz, P., Unser, M., and Thevenaz, P. (2000). Optimization of mutual information for multiresolution image registration. IEEE Trans. Image Process. *9*, 2083–2099.

Thompson, L.M. (2008). Neurodegeneration: A question of balance. Nature *452*, 707–708.

Thompson, P.M., Hayashi, K.M., de Zubicaray, G., Janke, A.L., Rose, S.E., Semple, J., Herman, D., Hong, M.S., Dittmer, S.S., Doddrell, D.M., et al. (2003). Dynamics of gray matter loss in Alzheimer's disease. J. Neurosci. *23*, 994–1005.

Trzepacz, P.T., Saykin, A., Yu, P., and Bhamditipati, P. (2014). NIH Public Access. *21*, 607–622.

Ueckert, S., Plan, E.L., Ito, K., Karlsson, M.O., Corrigan, B., Hooker, A.C., and Alzheimer's Disease Neuroimaging Initiative (2014). Improved Utilization of ADAS-Cog Assessment Data Through Item Response Theory Based Pharmacometric Modeling. Pharm. Res. *31*, 2152–2165.

Verma, N., Beretvas, S.N., Pascual, B., Masdeu, J.C., Markey, M.K., Rosen, W., Mohs, R., Davis, K., Cano, S., Posner, H., et al. (2015). New scoring methodology improves the sensitivity of the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog) in clinical trials. Alzheimers. Res. Ther. *7*, 64.

Waldemar, G., Dubois, B., Emre, M., Georges, J., McKeith, I.G., Rossor, M., Scheltens, P., Tariska, P., Winblad, B., and EFNS (2007). Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. Eur. J. Neurol. *14*, e1–e26.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing With Short Test Length. Educ. Psychol. Meas. *73*, 1017–1035.

Wang, J., Logovinsky, V., Hendrix, S.B., Stanworth, S.H., Perdomo, C., Xu, L., Dhadda, S., Do, I., Rabe, M., Luthman, J., et al. (2016). ADCOMS: a composite clinical outcome for prodromal Alzheimer's disease trials. J. Neurol. Neurosurg. Psychiatry jnnp-2015-312383.

Weiner, M.W., Aisen, P.S., Jack, C.R., Jagust, W.J., Trojanowski, J.Q., Shaw, L., Saykin, A.J., Morris, J.C., Cairns, N., Beckett, L.A., et al. (2010). The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimers. Dement. *6*, 202–211.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al. (2012). The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's Dement. *8*, S1--S68.

Welge, V., Fiege, O., Lewczuk, P., Mollenhauer, B., Esselmann, H., Klafki, H.-W., Wolf, S., Trenkwalder, C., Otto, M., Kornhuber, J., et al. (2009). Combined CSF tau, p-tau181 and amyloid-β 38/40/42 for diagnosing Alzheimer's disease. J. Neural Transm. *116*, 203–212.

Westover, M.B., and O'Sullivan, J.A. (2008). Achievable rates for pattern recognition. IEEE Trans. Inf. Theory *54*, 299–320.

Weyer, G., Erzigkeit, H., Kanowski, S., Ihl, R., and Hadler, D. (1997). Alzheimer's Disease Assessment Scale: reliability and validity in a multicenter clinical trial. Int. Psychogeriatrics *9*,

123–138.

Wilson, R.S., Barral, S., Lee, J.H., Leurgans, S.E., Foroud, T.M., Sweet, R.A., Graff-Radford, N., Bird, T.D., Mayeux, R., and Bennett, D.A. (2011). Heritability of different forms of memory in the Late Onset Alzheimer's Disease Family Study. J. Alzheimers. Dis. *23*, 249–255.

Woods, A., and Baker, R. (1985). Item response theory. Lang. Test. *2*, 117–140.

Wouters, H., van Gool, W.A., Schmand, B., and Lindeboom, R. (2008). Revising the ADAS-cog for a more accurate assessment of cognitive impairment. Alzheimer Dis. Assoc. Disord. *22*, 236–244.

Yum, S.W., Zhang, J., Mo, K., Li, J., and Scherer, S.S. (2009). A novel recessive Nefl mutation causes a severe, early-onset axonal neuropathy. Ann. Neurol. *66*, 759–770.