ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Validity and Reliability of Digital Biomarkers for Frontotemporal Dementia

by

IOANNIS STYLIDIS

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
February 2025

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Εγκυρότητα και Αξιοπιστία των Ψηφιακών Βιοδεικτών για τη Μετωποκροταφική Άνοια

ΙΩΑΝΝΗΣ ΣΤΥΛΙΔΗΣ

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα
Φεβρουάριος 2025

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisors, Dr. Ioannis Ntzoufras, Dr. John Kornak, and Dr. Alessandro Spelta, for their enormous support throughout this journey. I had also the opportunity to meet and collaborate with many thoughtful people while studying in Master's program in Statistics, from whom I learned so much. Lastly, I want to thank my family and friends for supporting me and being there when I needed them the most.

ABSTRACT

IOANNIS STYLIDIS

Validity and Reliability of Digital Biomarkers for Frontotemporal Dementia

February 2025

Frontotemporal dementia is a neurological disease that has been a concern for humanity for a long time. Patients with this disease suffer from various consequences, such as cognitive dysfunction, motor problems, and emotional dysregulation. Unfortunately, there is no permanent treatment for it. Therefore, finding ways to collect more data would be highly beneficial in addressing this problem. In response, a mobile application (ALLFTD Mobile App) has been developed that its purpose is to simulate the data collected from diagnostic tests. This application includes mobile games designed to assess the cognitive function of patients, which can be an insight information about the level of dementia in them. By gathering more data from patients, this application facilitates research efforts aimed at finding a treatment and making better predictions of the severity of the disease in patients.

In our analysis, we will demonstrate that this mobile application has the potential to replace certain diagnostic tests that are typically performed in hospitals, allowing patients to undergo fewer procedures. This will be achieved by using data from games' performances over time from participants and evaluating their association with diagnostic tests such as volumes of brain regions from MRI scans. In this study, two techniques-Mixed-effects models and Structural Equation modeling(SEM)-will be used to analyze the relationship between game performance and brain volumes.

ΠΕΡΙΛΗΨΗ

ΙΩΑΝΝΗΣ ΣΤΥΛΙΔΗΣ

Εγκυρότητα και Αξιοπιστία των Ψηφιακών Βιοδεικτών για τη Μετωποκροταφική Άνοια

Φεβρουάριος 2025

Η μετωποκροταφική άνοια είναι μια νευρολογική πάθηση που απασχολεί την ανθρωπότητα για μεγάλο χρονικό διάστημα. Οι ασθενείς με αυτή την πάθηση υποφέρουν από διάφορες συνέπειες, όπως γνωστική δυσλειτουργία, κινητικά προβλήματα και συναισθηματική αναστάτωση. Δυστυχώς, δεν υπάρχει μόνιμη θεραπεία για αυτήν. Επομένως, η αναζήτηση τρόπων για τη συλλογή περισσότερων δεδομένων θα ήταν εξαιρετικά χρήσιμη στην αντιμετώπιση αυτού του προβλήματος. Ως απάντηση, έχει αναπτυχθεί μια εφαρμογή κινητού (*ALLFTDMobileApp*), η οποία έχει ως στόχο να προσομοιώσει τα δεδομένα που συλλέγονται από διαγνωστικά τεστ. Αυτή η εφαρμογή περιλαμβάνει παιχνίδια κινητού σχεδιασμένα για να αξιολογούν τη γνωστική λειτουργία των ασθενών, τα οποία μπορεί να παρέχουν πληροφορίες για το επίπεδο της άνοιας στους ασθενείς. Συλλέγοντας περισσότερα δεδομένα από τους ασθενείς, η εφαρμογή αυτή διευκολύνει τις ερευνητικές προσπάθειες που αποσκοπούν στην εύρεση θεραπείας και στην καλύτερη πρόβλεψη της σοβαρότητας της νόσου στους ασθενείς.

Στην ανάλυσή μας, θα αποδείξουμε ότι αυτή η εφαρμογή κινητού έχει τη δυνατότητα να αντικαταστήσει ορισμένα διαγνωστικά τεστ που συνήθως πραγματοποιούνται στα νοσοκομεία, επιτρέποντας στους ασθενείς να υποβληθούν σε λιγότερες διαδικασίες. Αυτό θα επιτευχθεί χρησιμοποιώντας δεδομένα από τις επιδόσεις των παιχνιδιών με την πάροδο του χρόνου από τους συμμετέχοντες και αξιολογώντας τη συσχέτισή τους με διαγνωστικά τεστ, όπως οι όγκοι περιοχών του εγκεφάλου από εγκεφαλικές απεικονίσεις. Στη μελέτη αυτή, θα χρησιμοποιηθούν δύο τεχνικές — Μοντέλα Μικτών Επιδράσεων και Μοντελοποίηση Δομικών Εξισώσεων — για να αναλυθεί η σχέση μεταξύ των επιδόσεων στα παιχνίδια και των όγκων του εγκεφάλου.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Frontotemporal Dementia

The suffering that patients with frontotemporal dementia endure is immense. Having this disease makes every simple, easy task of the day seem difficult or even impossible to accomplish. They forget things, social interactions become frustrating for both themselves and others, language problems arise, and even control over their own bodies is diminished. Consequently, some essential skills required to interact with others are difficult to have. Empathy, emotional control and regulation are among these traits. Dementia with such symptoms is referred to as Behavioral Variant FTD.[5]

As a result, patients with these symptoms experience difficulties in their relationships because they cannot behave appropriately. They also face challenges in managing their financial decisions due to impulsive behavior, and overall, life becomes more difficult in every aspect, as emotions play a central role in daily functioning. Another subtype of frontotemporal dementia is Primary Progressive Aphasia. This subtype is further divided into progressive non-fluent aphasia and semantic dementia. The former affects patients by making proper and fluent speech difficult for them. In contrast, patients with the latter subtype have difficulty understanding the meanings of words. Table 1.1 presents all types of frontotemporal dementia, their symptoms, and the regions of the brain most affected.

Subtype	Key Features	Affected Brain Areas
Behavioral Variant FTD (bvFTD)	Personality changes, disinhibition, apathy, compulsive behaviors, emotional blunting	Frontal lobes (particularly the prefrontal cortex)
Nonfluent Variant PPA (nfvPPA)	Effortful, halting speech, agrammatism, apraxia of speech, preserved word meaning	Left posterior frontal lobe and insula
Semantic Variant PPA (svPPA)	Fluent but vague speech, loss of word meaning, impaired recognition of faces/objects	Anterior temporal lobes (left-sided)
FTD with Motor Neuron Disease (FTD-MND)	Motor symptoms (weakness, fasciculations), behavioral or language symptoms	Frontal and temporal lobes with motor areas
Corticobasal Syndrome (CBS)	Motor symptoms (stiffness, dystonia), apraxia, cognitive decline	Frontal and parietal lobes
Progressive Supranuclear Palsy (PSP)	Impaired vertical eye movements, postural instability, behavioral changes	Brainstem and frontal lobes

Table 1.1: Subtypes of Frontotemporal Dementia and their characteristics

Why does frontotemporal dementia occur? This happens because frontotemporal dementia is a type of brain disorder that primarily causes the frontal and temporal lobes to atrophy and shrink. While the full story of why and how this occurs remains unclear, there is evidence that genetic factors and certain protein abnormalities are linked to the physical deterioration of these brain regions. Below is some information about the genetic and protein abnormalities associated with the condition.

Genetic Factors and Protein Abnormalities

Approximately 30–50% of patients with frontotemporal dementia have specific genetic mutations associated with the disease. Mutations in genes such as MAPT (microtubule-associated protein tau), GRN (progranulin), and C9orf72 are known to be linked to FTD. These genetic mutations disrupt normal protein function, leading to neuronal cell death in the brain.[11]

Frontotemporal dementia patients appear to share a common underlying issue: certain proteins in brain cells do not function properly. Two well-known proteins associated with these diseases are tau and TDP-43. The abnormal accumulation of these proteins in cells causes neuronal degeneration, leading to a gradual reduction in the size of brain regions due to neuron death. For example, mutations in the MAPT gene on chromosome 17, which encodes the tau protein, result in the abnormal aggregation of this protein.[11]

However, as mentioned earlier, only 30–50% of patients worldwide develop this disease due to genetic mutations. For the remaining cases, there is no proven cause. Aging seems to be the most likely factor, as aging reduces cellular efficiency and increases the likelihood of protein misfolding. Additionally, environmental factors such as head injuries or chronic exposure to toxins may contribute to the disease. Lifestyle factors also play a significant role in cell health, meaning individuals with chronic stress, poor sleep, diabetes, or high cholesterol may be at higher risk of developing frontotemporal dementia. As a result, symptoms typically begin to appear between the ages of 45 and 65, and the disease accounts for 10–20% of all dementia cases.[5] This underscores the importance of early diagnosis and providing patients with the available treatments.

This study, along with the mobile application developed by researchers at the University of California, San Francisco, aims to address this need. The mobile games included in the application are designed to assess the stage of the disease in patients. In addition, the application facilitates the research for improved temporary treatments and potentially a permanent cure. However, the use of digital tests is not yet widely accepted and other diagnostic methods are currently preferred.

In Table 1.2, the most commonly used diagnostic methods are presented. These include MRI scans, clinical evaluations, neuropsychological tests, biomarkers, and genetic testing.

Diagnostic Method	**Purpose**
Clinical Assessment	Identify behavioral, language, or motor symptoms.
Neuropsychological Testing	Evaluate cognitive domains and distinguish from other dementias.
MRI / CT	Detect frontal/temporal lobe atrophy.
PET / SPECT	Assess reduced metabolism in affected brain regions.
Biomarkers (CSF, blood)	Differentiate from Alzheimer's; genetic testing for familial cases.
Speech & Language Testing	Diagnose specific variants of Primary Progressive Aphasia (PPA).
Genetic Testing	Identify mutations in familial cases.

Table 1.2: Diagnostic Methods

1.2 Motivation

Diagnostic tests are used to assess the stage of the disease, after which treatments are available to help manage symptoms. However, a permanent cure has yet to be discovered. Some of the temporary treatments are listed in Table 1.3, with the most common approaches including medications and therapies such as occupational therapy, physiotherapy, and social interaction therapy. Developing more effective treatments requires further data collection.

Currently, available research data primarily consist of biomarkers from MRI scans, biofluid measurements, and genetic analysis. However, collecting these biomarkers is costly, time-consuming, and procedurally complex. This is where the serious games developed by the University of California, San Francisco, become a valuable asset in the search for a permanent cure. These games allow data to be collected through smartphones, making the process more accessible and efficient. Given the widespread use and acceptance of smartphones today, this approach significantly simplifies data collection and increases its volume.[2] The key challenge lies in determining whether the data obtained from these games are sufficient for conducting robust statistical analyses and deriving meaningful insights. Further research is needed to validate their effectiveness as reliable biomarkers for tracking disease progression.

Treatment Category	**Desctription**
Medications	Antidepressants (e.g., SSRIs): Offsetting the hormones, so the behavioral symptoms will be lighter.
Therapies	Therapies for behavior, speech and language
Lifestyle	Optimization of diet, exercise, emotional regulation and sleep

Table 1.3: Treatments for Frontotemporal Dementia

1.3 Literature

Researchers have been investigating the validity of digital biomarkers as tools for assessing cognitive function and detecting neurodegenerative diseases. Over the years, many games have designed for this particular purpose. One of the most well-known examples is Sea Hero Quest, a mobile game specifically developed to evaluate spatial navigation deficits—one of the earliest signs of Alzheimer's disease. This game has generated a vast amount of data, allowing researchers to uncover important insights into age-related and gender-related differences in spatial navigation abilities, as well as other cognitive functions.[8]

A particularly relevant study for our research examined the relationship between game performance and established dementia biomarkers, using digital cognitive assessments as predictors and other biological and clinical measures as outcomes.[2] The study found that the internal consistency of these game-based cognitive tests, ranged from 0.77 to 0.95 (Cronbach's alpha coefficient) , which indicated good to excellent reliability. Additionally, intraclass correlation (ICC) was used for test-retest reliability and the results are between 0.77 and 0.95, indicating moderate to high stability across multiple test sessions. Another valuable result is the diagnostic performance of the games for distinguishing symptomatic individuals from healthy controls. Using the Area Under the Curve(AUC) as a method for this measurement, they concluded that games are highly effective in identifying cognitive impairment (AUC=0.93).Notably, the findings also revealed that these assessments were more sensitive to early symptoms than some traditional cognitive screening measures. It is important to mention the validity of these digital assessments in relation to other biomarkers of dementia because it is the main goal of our study. The study reported significant correlations between game performance and clinical and neuroimaging measures, including:

CDR®+NACC-FTLD scores (r = 0.38–0.59): Widely recognized as a key biomarker for frontotemporal dementia.

Neuropsychological tests scores (r = 0.40–0.66): Further validating the effectiveness of game-based assessments in capturing cognitive function.

Brain volume measurements (r = 0.34–0.50): Reinforces the potential of game-based cognitive tests as digital biomarkers for neurodegeneration.

In 2021, JMIR Serious Games contributed to this topic by assessing whether digital card games could be used as biomarkers for mild cognitive impairment. Mild cognitive impairment (MCI) is strongly associated with Altzheimer's and frontotemporal dementia since patients with MCI are more likely to develop it. Most commong way to dianose MCI is by using standardized cognitive tests, such as the Montreal Cognitive Assessment. However, in this study they investigated whether performance in a digital card game called Klondike Solitaire, could evaluate the health statement of participants. Participants are separated to healthy individuals and people diagnosed with MCI. They played the game and the aim was to see if there were differences between healthy and unhealthy individuals. To analyze the data, they performed Generalized Linear Mixed Model(GLMM) with response variable measurements of the game. As fixed effects they included a categorical variable that indicates if the patient is healthy or has the disease. In addition, they controlled for age, tablet proficiency and Klondike Solitaire experience.

The last two variables ensure that results were not due to lack of familiarity. By that model, it is possible to see if participants with the disease have worse results in the game compared to the healthy participants. The analysis revealed that 12 out of 23 gameplay-related measures there were statistically significant differences between healthy controls and individuals with MCI. These findings could be useful for detecting early signs of cognitive decline.[12]

Another important study explored the use of a game-based approach to identify digital biomarkers for Parkinson's Disease (PD), leveraging a mobile application to collect detailed motor and cognitive data in an engaging and accessible format. The research design involved a cross-sectional comparison between individuals diagnosed with PD and a control group of healthy participants, matched by age. Participants interacted with a series of touchscreen-based tasks designed to evaluate fine motor skills (e.g., tapping, swiping, and tracing), reaction time, and cognitive functions such as memory and decision-making. Data were gathered through the device's built-in sensors, including touchscreen inputs, accelerometers, and gyroscopes, which captured precise kinematic and temporal metrics. Advanced machine learning techniques were applied to analyze the data, with feature extraction and selection processes used to identify patterns associated with PD. Classification models, such as random forest and support vector machines, were trained to differentiate between PD patients and controls based on the extracted features. The results highlighted several key digital biomarkers, including increased variability in tapping speed, reduced precision in tracing tasks, and slower reaction times, which were strongly associated with PD. The models achieved an overall accuracy of over 85%, with sensitivity and specificity rates above 80%, indicating high diagnostic reliability. Furthermore, cognitive tasks involving memory and timed decision-making were found to provide additional diagnostic insights. These findings demonstrate the potential of game-based digital tools as a scalable, non-invasive method for early PD detection and monitoring, offering a promising complement to traditional clinical evaluations.[1]

In the article called "Concurrent Validity, Test-Retest Reliability, and Normative Properties of the Ignite App: A Cognitive Assessment for Frontotemporal Dementia“, researchers evaluated the Ignite app's effectiveness as a digital cognitive assessment tool for Frontotemporal dementia (FTD). The study comprised two main cohorts: a normative group of over 2,000 cognitively healthy adults aged 20 to 80 years, recruited remotely to establish baseline performance metrics, and a secondary cohort of 98 healthy controls who completed the Ignite app twice, seven days apart, alongside traditional pen-and-paper neuropsychological tests and a user experience questionnaire. The app assessed multiple cognitive domains, including executive function, processing speed, and social cognition. Results indicated significant associations between age and performance on the app's processing speed (r = 0.42–0.56) and executive function tasks (r = 0.43–0.62), highlighting the app's sensitivity to age-related cognitive changes. Test-retest reliability analyses demonstrated moderate to excellent stability, with intraclass correlation coefficients ranging from 0.54 to 0.92 across various tasks. Additionally, the app's measures showed significant correlations with traditional neuropsychological tests (r = 0.25–0.72), supporting its concurrent validity. User feedback was overwhelmingly positive, with over 90% of participants finding the app enjoyable and easy to use without supervision. These findings suggest that the Ignite app is a valid, reliable, and user-friendly tool for assessing cognitive function in FTD, offering a scalable solution for remote cognitive assessments in both clinical and research settings.[7]

In the pilot study titled "Using Self-Administered Game-Based Cognitive Assessment to Screen for Degenerative Dementia", researchers developed and validated a Game-Based Cognitive Assessment (GBCA) tool aimed at early detection of degenerative dementia in older adults. The study involved 67 patients diagnosed with neurocognitive disorders and 57 healthy controls. Participants completed the self-administered GBCA alongside traditional cognitive assessments, including the Clinical Dementia Rating (CDR), Cognitive Abilities Screening Instrument (CASI), and Mini-Mental State Examination (MMSE). Additionally, a user-experience questionnaire was administered to evaluate the tool's usability and acceptability. Statistical analyses revealed that the average GBCA scores were significantly higher in healthy controls (87 ± 7.9) compared to the NCD group (52 ± 21.7), indicating the tool's effectiveness in distinguishing between the two groups. The GBCA demonstrated strong concurrent validity, correlating well with CASI ($r^2 = 0.90$, $p < .01$) and MMSE ($r^2 = 0.92$, $p < .01$) scores. A GBCA cut-off score of 75/76 yielded a sensitivity of 85.1%, specificity of 91.5%, and an area under the curve (AUC) of 0.978, with positive and negative predictive values of 91.9% and 84.4%, respectively. User-experience feedback indicated that both healthy controls and NCD participants found the GBCA acceptable and user-friendly, with particularly positive responses from the healthy control group. These findings suggest that the GBCA is an effective and user-friendly tool for screening degenerative dementia.[22]

1.4 Contribution

Although the literature on this topic is extensive, our study approaches the problem from a different perspective. Machine learning techniques are highly effective in predicting and monitoring the progression of dementia, but they provide limited insight into why dementia occurs. Understanding the underlying causes of brain volume deterioration is crucial for developing a permanent cure.

Previous studies, as discussed in the earlier section, have primarily focused on cross-sectional data, which capture only a single snapshot of a patient's condition. This limitation makes it difficult to track changes over time or establish cause-and-effect relationships. For instance, if a study finds that patients with dementia experience poorer sleep, it remains unclear whether poor sleep contributes to dementia or if dementia causes sleep disturbances. In contrast, longitudinal datasets allow for continuous monitoring of dementia progression, enabling more reliable conclusions. Additionally, repeated measurements over time can help track how individual patients respond to treatments, distinguishing disease progression from natural variations between individuals. This study aims to leverage these advantages of longitudinal data.

As mentioned earlier, collecting brain MRI scans and biofluid data is particularly challenging, especially for patients with frontotemporal dementia and when repeated measurements are required. Therefore, assessing the relationship between brain volumes from MRI scans and mobile game performance over time could provide a valuable alternative. If meaningful patterns of brain atrophy can be identified through game performance, these mobile games could serve as a proxy for repeated measurements, allowing for earlier and more accurate disease assessments.

Longitudinal data collected through these games would also offer significant benefits for research. Early diagnosis of dementia would become more feasible, and treatments could be administered at earlier stages, improving their effectiveness. Furthermore, if these games prove reliable in clinical trials, tracking disease progression would become more practical, making the search for a cure more attainable.

In this study, we use several mobile games developed by the ALLFTD research team to evaluate their effectiveness in predicting changes in brain region volumes. The size of these brain regions provides critical information about dementia progression, but obtaining these measurements regularly through MRI scans is difficult. Using game performance as an alternative biomarker would simplify this process. To analyze the data, we employ mixed-effects models and structural equation modeling, which are particularly well-suited for handling longitudinal data with repeated measurements.

In the next chapter, we will review the theoretical framework of longitudinal analysis and the models applied in this study. Specifically, we will discuss data-related challenges and potential solutions. We will then introduce mixed-effects models and the bivariate latent change score model to examine the relationship between game performance and brain region volumes over time. Finally, we will apply these models to our dataset to evaluate the validity and reliability of ALLFTD mobile games as digital biomarkers over time.

Chapter 2

Methodology

2.1 Longitudinal analysis

When working with repeated measurements over time, longitudinal analysis is essential as it captures both within-subject and between-subject variability. In cross-sectional studies, each subject provides only a single snapshot of data, allowing for the analysis of differences between individuals but offering no insight into individual trajectories over time. In contrast, longitudinal studies track the same individuals across multiple time points, enabling the observation of changes over time and individual differences in those trajectories. Additionally, this approach allows us to examine the factors that influence these variations.[10]

In this study, longitudinal analysis is used to assess the relationship between brain volumes and game performance over time. More specifically, we investigate whether changes in game performance provide meaningful insights into changes in brain volume. For instance, participants who achieve higher scores in the games may exhibit different patterns of brain volume change over time. By modeling these relationships, we can gain a deeper understanding of cognitive function progression and neurodegeneration.

Traditional statistical methods, such as standard regression models and t-tests, are inadequate for analyzing longitudinal data as they do not account for the correlations inherent in repeated measurements.[18] As a result, they often yield unreliable and biased estimates. Longitudinal analysis, on the other hand, is specifically designed to handle such data structures, making it ideal for studying both within-subject changes and differences between individuals.

To analyze our data, we employ mixed-effects models and structural equation modeling, both of which are well-suited for handling repeated measurements. Each method has distinct advantages that contribute to obtaining valid and reliable results. In the following sections, we will describe these approaches in detail and outline their respective benefits.

2.2 Missing Values

The presence of missing values introduces significant challenges in the analysis. The most critical issue is the loss of valuable information, which can lead to less precise estimates and unreliable model coefficients.[3] In addition, missing data can inflate standard errors, weakening the statistical power of hypothesis tests and reducing the accuracy of parameter significance assessments. To improve inference and obtain more reliable results, it is crucial to address the issue of missing data effectively. Missing values can arise due to three primary mechanisms.[3]

1. **Missing completely at random:** The missingness occurs independent of the observed and unobserved data. An example is when accidentally a participant skips a question in a survey. In that situation, missing values do not provoke any problem in the analysis and the coefficients remain unbiased.

2. **Missing at random:** Missingness dependes on observed data and not on the unobserved values. For instance, in a illness someone could not have participate in the end of the survey because of severeness of the disease and inability to do some particular things. If this will not be assesed right in analysis and missing values will be solved like they are missing completely at random then biased will occur in the estimates.

3. **Missing Not at random:** Missingness depends on unobserved data themselves. In that case, more information is required about the reason of missingness and a more complicated techniques to resolve them.

There are several approaches to addressing the issue of missing values. The simplest method is complete case analysis, which involves discarding all data points with missing values. This approach is valid under the assumption of Missing Completely at Random (MCAR), meaning the missingness is unrelated to any observed or unobserved data. However, if this assumption does not hold, complete case analysis can introduce bias and increase the variability of estimates due to the loss of valuable information.[3] A more flexible approach is available data analysis, which utilizes all available observations without discarding entire cases. While it retains more data than complete case analysis, it may still be suboptimal compared to more sophisticated techniques.

Advanced methods such as multiple imputation and maximum likelihood estimation (MLE) are commonly used under the assumption of Missing at Random (MAR). These techniques provide more accurate and less biased estimates compared to simpler methods. Multiple imputation replaces missing values with several plausible estimates, creating multiple datasets that are analyzed separately and combined for inference. MLE, on the other hand, estimates parameters by maximizing the likelihood function, making full use of all available data without imputing missing values explicitly.[21]

For cases where data are Missing Not at Random (MNAR), more complex statistical models are required. Joint modeling strategies estimate both the distribution of the observed variable and the missing data mechanism simultaneously, making them suitable for handling MNAR data.[21]

In this study, we primarily use maximum likelihood estimation, as our modeling approach naturally incorporates this method to estimate coefficients efficiently and accurately. Additionally, an interpolation technique will be applied to impute certain missing values, which will be further detailed in the Results section.

While simpler imputation techniques exist, such as mean imputation or regression imputation, they can introduce bias and distort statistical inference by underestimating variability and leading to incorrect conclusions. Therefore, these methods should be avoided whenever possible in favor of more robust approaches.[21]

Linear Interpolation

As we will see in the next chapter, a key complication in our dataset limits the use of certain statistical tools for analyzing the longitudinal relationship between game performance and brain volumes. Specifically, while game performance is measured every six months, brain volumes from MRI scans are recorded only once per year. As a result, we have at most two time points for brain volume measurements, making it impossible to fit a random slope in mixed-effects models.

To address this issue, we will apply linear interpolation to impute missing values between the two MRI measurements. This approach allows us to estimate intermediate brain volume values, enabling a more comprehensive longitudinal analysis of the relationship between game performance and brain atrophy over time. The interpolation formula is the following:

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0}$$

where,

- (x_0,y_0) and $(x_1,\, y_1)$ known points

- x_0 and x_1 represent the time points

- y_0 and y_1 represent the values of brain volumes

Linear interpolation constructs a straight-line using the two known points of brain volumes and derive new data points within the range of them. It assumes a constant rate of change between the two brain volume measurements.

2.3 Mixed-Effects Models

Mixed-effects models, often termed hierarchical or multilevel models, are statistical tools designed to account for variability at multiple levels within a dataset. These models are particularly useful for analyzing data that exhibit a nested or clustered structure, such as repeated observations within individuals or groups. More specifically, we build a mixed-effects model when there is a need to analyze the variability of within and between-subjects effects. Between-subject effects are variables that do not change over time. If we lack information about these kinds of data, participants may exhibit different baseline levels. For instance, gender is this kind of variable and if we do not have this information there will be an unexplained error if we use another kind of modeling. Mixed effects models can capture this variability, thereby improving the precision of inference. Furthermore, variability between subjects can provide valuable insight into factors that are missing as fixed effects and may help identify variables that explain differences between participants. On the contrary, within-subject effects are those that change over time. These effects capture the variability of the response variable within an individual across different time points or conditions.[16]

Other approaches of modeling those kind of data are not particularly suitable. For example, generalized linear models do not include random effects, so they can not capture the within-subject variability. Also, this correlation that presents in the within-subject observations violates the assumption of independence. In addition, GLM do not include random intercepts in order to capture the subject-specific variability, thus there will be an unexplained overdispersion. These issues results to incorrect inferences. Mixed effects models allow for subject-specific variability through random effects, accommodating both within-subject correlations and unbalanced data structures. This flexibility ensures a more accurate representation of the hierarchical nature of the data, such as repeated measurements. In conclusion, mixed effects models are particularly useful if the purpose of the researchers is also to estimate the within-subject variability in order to decide about the future of the patients. For example, if an individual has severe symptoms, these kind of models can help to identify it and act appropriately for future care.

Distinguishing Fixed and Random Effects

In mixed effects models, effects can be categorized as fixed or random:[10]

- **Fixed Effects:** Represent consistent, population-wide impacts of predictors. These effects are assumed to be the same across all groups or individuals in the dataset.
- **Random Effects:** Capture variability specific to groups, clusters, or individuals within the dataset. These effects allow for group-specific deviations from the overall population trends.

Random effects are especially useful for modeling situations where variability exists due to grouping factors, such as repeated measurements from the same subject. This could represent differences in baseline brain region volumes across subjects.

Two common components of random effects are:

- **Random Intercepts:** Allow groups or individuals to have different baseline levels of the outcome variable. For example, subjects might have varying average brain region volumes regardless of their mobile game scores, age, or sex.
- **Random Slopes:** Enable the relationship between a predictor and the outcome to vary across groups or individuals. For instance, the relationship between mobile game scores and brain region volumes might differ among subjects, reflecting individual-specific trends.

By incorporating random intercepts and slopes, mixed effects models can provide a more nuanced understanding of variability in your data. Figure 2.1 presents four different models with different assumptions regarding random effects. The first plot, which demonstrates a simple linear regression model, assumes homogeneity in the data, which means that all participants share the same intervept and slope. In contrast, the mixed-effects model with only a random intercept assumes for each group a different baseline value but the same fixed slope. For example, each participant in our study could have a different baseline level of brain volumes, but the trajectory of brain volumes over game performance remains the same. Another assumption for the random effects could be the fixed intercept with random slopes which means that participants have the same baseline, but the trajectories differ over time. And lastly, we can assume that everything is random for every subject, as in the last plot in Figure 2.1.

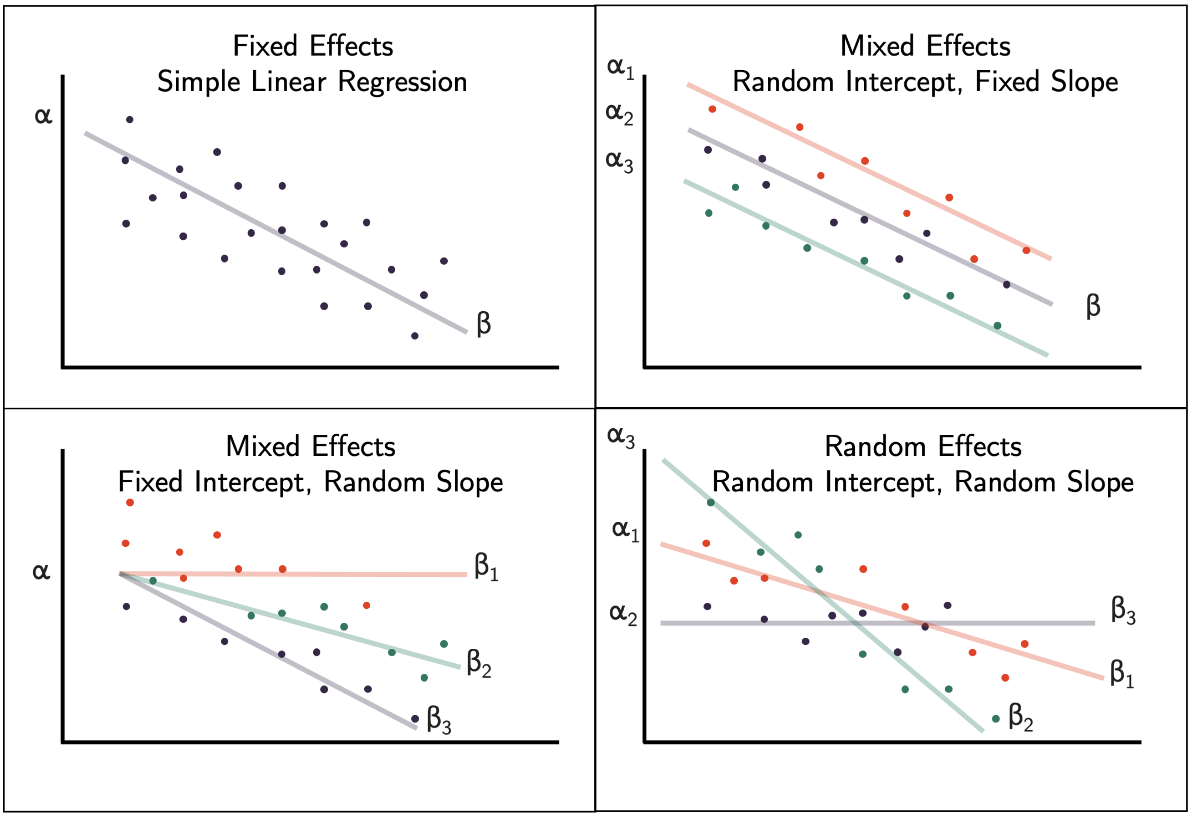


Figure 2.1: Different assumptions of random effects in mixed-effects models

Mathematical Representation

The mixed effects model that we will use in our analysis can be expressed as:

$$Y_{ij} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Time}_{ij} + \beta_3 x_{ij} + b_{0j} + b_{1j} x_{ij} + \epsilon_{ij}, \tag{2.1}$$

where:

- Y_{ij} denotes the brain region volume for subject j at time i,
- x_{ij} represents the mobile game score for subject j at time i,
- $Time_{ij}$ represents the semester i for subject j for $i = 0, 1, 2, ..., n_j$,
- β_0, β_1, β_2, and β_3 are fixed effect coefficients (intercept, effect of baseline age, effect of time, and effect of game scores, respectively),
- $b_{0j} \sim N(0, \sigma^2_{b_0})$ represents the random intercept for subject j,

- $b_{1j} \sim N(0, \sigma^2_{b_1})$ represents the random slope for the relationship between game scores and brain volume for subject j,
- $\epsilon_{ij} \sim N(0, \sigma^2)$ denotes residual errors.

This formulation accounts for variability in baseline brain region volumes and individual-specific relationships between game scores and brain region volumes while incorporating fixed effects for baseline age and time. The b_{0j} and b_{1j} captures the variability that is presented between-subjects. More specifically, b_{0j} detects the variability that exists in the baseline of subjects and the b_{1j} detects the variability that exists in the slopes of the subjects. Therefore, if there is a difference in the trends of the subjects, then this coefficient has the ability to fit a slope separately for each individual that vary from the population slope. After accounting for these variabilities, the residual error ϵ_{ij} is responsible for the variance left over from each subject:

$$\epsilon_i \sim \mathcal{N}_{n_i}(0, R_i).$$

Different Mathematical Representation

Another way that this model is presented is by two stages model. In the first stage, a model for the j subject , $j = 1, 2, ..., m$ is fitted:
Level 1 Within-Subject Model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \beta_{2j} \text{Age}_j + \beta_{3j} \text{Time}_{ij} + \epsilon_{ij}$$

This model fits a unique straight line for each subject. The deviations of subject's observations from this line is the ϵ_{ij} measurement errors. Each participant has coefficients $\beta_j = \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix}$.

However, the above model does not tell us anything about the among-subject variation. So, what about the population averages and the overall association of the variables with the outcome? Therefore, the above models need an adjustment in order to contain this information. Let's suppose that every participant is a part of a broader population. From the previous model, unique intercepts and slopes were constructed for each participant. Thus, we can describe the population as random vectors β_i. These random vectors distinguish one subject's trajectory from another.

Thinking About the Population

- It's helpful to imagine this population as being centered around an average intercept and slope. Some participants will have values closer to this average, while others may deviate with steeper or flatter slopes.
- More formally, we can describe the mean intercept and slope of the population using a central vector. Each subject's intercept and slope can then be seen as varying around this mean. To account for this variability, we can think of the population of β_i vectors as following: a joint probability distribution, which describes all possible values the random regression vector β_i could take.

Modeling the Population

To model this idea, let β_0 represent the average intercept and β_1 represent the average slope. Together, these can be written as:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Here, β represents the mean vector of the entire population of β_i. For any individual, we can then express their specific intercept and slope as:

$$\beta_i = \beta + b_i, \quad b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}.$$

This equation shows that each subject's β_i is made up of the population mean β plus a deviation vector b_i, the random effect, which captures what makes that subject's trajectory unique.

In a more formal sense, the vectors b_i are assumed to have a mean of 0 and are characterized by a covariance matrix. This matrix explains the variability in the intercepts and slopes between participants and how they are related. For example, if intercepts and slopes are increasing together, then that indicates a correlation between the random effects. The covariance matrix (D) captures these relationships. Moreover, the b_i vectors are assumed to follow a multivariate normal distribution that incorporates this covariance structure:

$$b_i \sim \mathcal{N}_k(0, D)$$

with $k = 2$ for a model with random intercepts and slopes.

Individual and Population-Level Models

The within-subject model and the population-level model play different roles in the explanation of the dataset. The first one describes the variation that is happening in the individual level and the latter one captures the variability in intercepts and slopes among all participants, linking the population-level and individual-level descriptions.

By combining the equations for β_{0i} and β_{1i} in the individual model, we can rewrite the overall model as:

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{ij} + e_{ij}.$$

This equation has the deviations of the intercepts and slopes of each subject (b_{0i}, b_{1i}) from the population averages (β_0, β_1). Therefore, this model has information for each participant separately and for the entire population. If there is a categorical variable like gender that distinguishes the population in different parts, then every group can have their own mean values of intercept and slope. For instance, for the gender, men could have different fixed parameters of mean intercept and slope than women.
If i subject is man:

$$\beta_{0i} = \beta_{0,M} + b_{0i}, \quad \beta_{1i} = \beta_{1,M} + b_{1i},$$

while if i is woman,

$$\beta_{0i} = \beta_{0,F} + b_{0i}, \quad \beta_{1i} = \beta_{1,F} + b_{1i}.$$

The way mixed effects model evaluate the parameters is not the same as the linear regression. The model exploits the nature of the coefficients and the multivariate normality assumption and uses the method of maximum likelihood. This works by estimating the parameter values in order to maximize the probability of a situation happens based on the data we observed. To understand the likelihood of observing the given data, we can describe it using the joint density function derived from the multivariate normal distribution. Assuming a multivariate normal model, the likelihood that the observed data vector y_i takes a specific value is captured by the joint density function for the multivariate normal distribution. Specifically, if we assume

$$Y_i \sim \mathcal{N}_{n_i}(X_i\beta, \Sigma_i),$$

then the probability of observing y_i is described as:

$$f_i(y_i) = (2\pi)^{-n_i/2}|\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - X_i\beta)^\top \Sigma_i^{-1}(y_i - X_i\beta)\right\}.$$

Since the observations Y_i are assumed to be independent, the joint density for all Y is simply the product of their individual densities. Letting $f(y)$ denote the joint density for the entire dataset Y, we have:

$$f(y) = \prod_{i=1}^{m} f_i(y_i) = \prod_{i=1}^{m} (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp\left\{ -\frac{1}{2}(y_i - X_i\beta)^\top \Sigma_i^{-1} (y_i - X_i\beta) \right\}.$$

Maximum Likelihood Estimators

The principle of maximum likelihood involves determining parameter values that maximize the likelihood function $f(y)$ for the observed data. In our model, these parameters include the fixed effects β and the covariance components ω (parameters of Σ_i). The values that maximize $f(y)$ are known as the maximum likelihood estimators, and they are functions of the observed data y.

However, the likelihood function depends on both β and ω, and it is generally complex. Specifically, solving for these parameters analytically is not feasible for most practical datasets due to the lack of a closed-form solution. Instead, numerical optimization techniques are employed to find the parameter values that maximize the likelihood.

Unknown Parameters: β and ω

In practical scenarios, it is realistic to assume that both β and ω are unknown. Under these circumstances, the maximum likelihood (ML) estimators must be derived for both parameters. The expressions for the estimators cannot be represented in neat closed forms. Instead, numerical algorithms are required to compute their values. This estimator for β is referred to as the generalized least squares (GLS) estimator. The term ”generalized” emphasizes that the covariance matrix Σ_i is not explicitly known and instead relies on its estimated values.

When ω is known, the exact sampling distribution of $\hat{\beta}$ can be derived, ensuring that it remains an unbiased estimator of β. However, when ω is unknown, the covariance matrices Σ_i are substituted by their estimates $\hat{\Sigma}_i$. As a result, it becomes infeasible to calculate the exact mean, covariance matrix, or other properties for $\hat{\beta}$. In such cases, asymptotic approximations are often employed for inference purposes.

Asymptotics of $\hat{\beta}$

The covariance matrix $\hat{\Sigma}_i$, which depends on $\hat{\omega}$, is itself a function of the data Y_i. This dependency makes it challenging to calculate properties of $\hat{\beta}$ in closed form. Consequently, $\hat{\beta}$ no longer has an exact p-variate normal sampling distribution. In such situations, deriving precise results becomes impractical. Instead, approximations are made under simplifying assumptions. A common approach is to consider what happens when the sample size becomes very large. Mathematically, this involves examining the behavior of $\hat{\beta}$ under the assumption:

$$m \to \infty,$$

where m is the number of units in the dataset.

Under these conditions, it can be shown that the sampling distribution of $\hat{\beta}$ is approximately unbiased. This means that $\hat{\beta}$ becomes an accurate estimator in the limit as m grows infinitely large. These results, however, are approximations. They describe what happens in an idealized scenario with an infinitely large sample size. For finite sample sizes, these results are only approximately valid. When m is moderately large, the approximations are generally quite accurate, though the exact threshold for m to be considered "large" is often unclear.

The results of large sample theory are fundamental to statistical methodology. For our case, they allow us to approximate the distribution of $\hat{\beta}$. Specifically, for large m, it can be shown that:

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \left(X^\top \hat{\Sigma}^{-1} X\right)^{-1}\right).$$

This approximation is a cornerstone of inference in linear mixed models, where exact distributions are otherwise difficult to derive. However, it is worth noting that any inference that are derived from these estimates should be taken with caution. This is the case because the assumptions that we made in order to derive these results may not be entirely true. In essence, the sampling size may not be large enough or the data do not follow the multivariate normal assumption.[10]

Restricted maximum likelihood

Although, maximum likelihood estimators provide unbiased coefficients β, it is different story for the parameters of the covariance matrix (ω). When the sample size is not too large, the estimates of covariance matrix are biased , providing inaccurate results for parameters of variances. Maximum likelihood estimates the parameters ω assuming that we have the true parameters of coefficients β. Thus, it does not account for the fact that β are estimated along with ω. Restricted maximum likelihood estimators, on the other hand, accounts for the estimated parameters β and adjusts the likelihood in order to deal with this issue. The resulting estimator for ω is less biased than the maximum likelihood estimators.[10]

Hypothesis testing and Diagnostics

After these assumptions that we made and the estimations of the parameters, we can make hypothesis about the significance of the variables. Therefore, we use standard error of coefficients and we build confidence intervals. Moreover, we can make some test statistics such as wald tests and likelihood ratio tests. It has been observed that likelihood ratio tests have better results when the sample is not too large.

Comparison Method	**Description**
Likelihood Ratio Test (LRT)	Compares two nested models by testing if the simpler model fits as well as the more complex model.
Akaike Information Criterion (AIC)	A measure of model fit that penalizes for model complexity. Lower AIC indicates a better model.
Bayesian Information Criterion (BIC)	Similar to AIC but with a stronger penalty for complexity. Lower BIC suggests a better model.
Marginal R-squared (R_m^2)	Explains the proportion of variance explained by fixed effects in the model.
Conditional R-squared (R_c^2)	Includes both fixed and random effects when calculating variance explained.
Residual Diagnostics	Analyzes residuals for patterns, such as heteroscedasticity or lack of normality, to assess fit.
Intraclass Correlation Coefficient (ICC)	Quantifies the proportion of variance explained by random effects in the model.
Fixed Effects Significance Tests	Tests the significance of individual fixed effects using Wald tests or t-tests.

Table 2.1: Common Methods for Comparing Mixed-Effects Models

2.4 Structural Equation Modeling

Structural Equation modeling(SEM) is a diverse set of methods that is widely used in the social and behavioral science fields. It is particularly useful when you want to combine confirmatory factor analysis and multiple regression. Essentially, using SEM we try to test some desirable hypothesis about associations and variability. This is achieved mainly by estimating means, variances and covariances of observed data. In addition, it is quite simple to make a structure that a response variable in one regression to be an explanatory variable in another equation. These models can be presented by using path diagram, which can help to understand intuitively the complex connections of the observed and latent variables. The main components of SEM are the following:

1. **Latent Variables:** Unobserved variables that are derived from observed indicators and usually they describe theoretical concepts that can not be measured directrly. For instance, in this study cognitive function will be used as a latent variable using games' scores as indicators.

2. **Observed Variables:** Directly measured indicators and used to derive the latent variables.

3. **Measurement Model:** The model that is used to present the latent variables and their observed indicators, which is included to confirmatory factor analysis.

4. **Structural Model:** This model presents the relationships among latent variables.

In Figure 2.2, there are the symbols that are used to describe the Structural Equations Model in a path diagram. These symbols will be used to describe the model that we will use in this study.

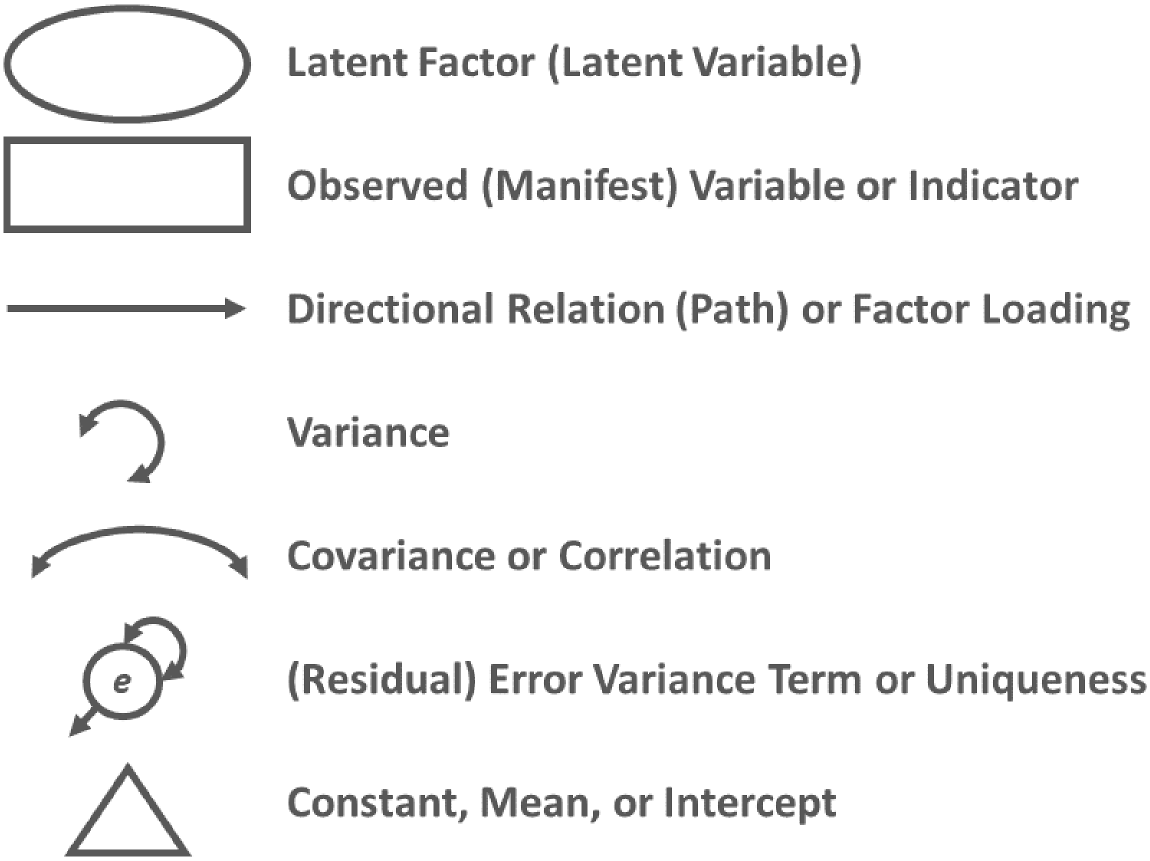

Latent Factor (Latent Variable)
Observed (Manifest) Variable or Indicator
Directional Relation (Path) or Factor Loading
Variance
Covariance or Correlation
e
(Residual) Error Variance Term or Uniqueness
Constant, Mean, or Intercept

Figure 2.2: Symbols in path diagram

Advantages of Structural Equation Modeling

There are four main reasons to use SEM:[23]

Validity: Using multiple indicators for latent variables enhances validity. By incorporating all cognitive games as indicators of a latent variable, we can extract meaningful information from each game while reducing the influence of unreliable or task-specific variance.

Reliability: Accounting for measurement error—caused by external factors and other influences—leads to more reliable and unbiased estimates, ensuring that the results reflect true underlying patterns rather than random fluctuations.

Complex models: Structural Equation Modeling (SEM) allows for the implementation of sophisticated models that would be significantly more challenging to estimate using traditional statistical approaches.

Hypothesis testing: SEM provides a framework for testing theoretical models by comparing them to actual data, allowing researchers to assess model fit and refine their assumptions based on empirical evidence.

Confirmatory Factor Analysis

All measurements are subject to error. Confirmatory Factor Analysis (CFA) is ideal for accounting for this measurement error. Our goal is to isolate the true score of the measurement to obtain more reliable results when associating them with other variables. Let's define X as the observed variables and as t the true value without the error. Thus, our objective is to decompose the true value from the error:[19]

$$X = t + \epsilon$$

Given that we analyze app cognitive games played by people with dementia - who often exhibit high variability in mood, attention, and behavior - it is beneficial to separate the true cognitive performance of the participants from these errors. The latent variable introduced through CFA aims to assess cognitive function and determine its association with brain volumes. Errors arising from mood fluctuations, distractions, inconsistencies, or other random factors can bias coefficient estimates. For example, a study by the ALLFTD research team concluded that phone-based performance assessments are quite reliable.[20] However, some individuals may lack familiarity with smartphones, especially older adults, introducing bias into the estimates. In addition, some participants can be more distracted than others due to factors that are not available to us. Moreover, since games measure slightly different aspects, the latent factor can capture their common variance. Various systematic errors can be accounted from Confirmatory Factor Analysis, and improving the reliability and validity of inferences.

Besides CFA, Exploratory Factor Analysis (EFA) has been used to address this problem. EFA defines factor loadings to best reproduce correlations between observed variables, constructing the same number of factors as variables. Subsequently, a smaller number of factors that explain a satisfactory amount of observed variance are retained. However, EFA has limitations. Firstly, it is an inductive procedure, meaning the data dictate the theory. Conversely, starting with a theory is generally more insightful, as it allows for hypothesis formulation and subsequent statistical testing against sample data. Therefore, an advantage of CFA is the ability to specify the measurement model before examining the data. We impose constraints between factors and items, which can then be tested for statistical significance. These parameter constraints can render the model over-identified, which is desirable for hypothesis testing. Constraints may include setting certain factor loadings to zero, defining the variance of the latent variable to one, or constraining one item's factor loading to one. The latter two constraints serve to define the metric of the latent variable, which generally lacks a predefined scale. By setting the variance of the latent variable to one, we obtain standardized solutions for the factor loadings, enhancing interpretability. Alternatively, constraining an item's factor loading to one assigns the latent variable the same scale as that item. For instance, if a game's score ranges from 0 to 10, the latent variable will adopt the same scale, allowing us to obtain unstandardized factor loadings.

Figure 2.3 illustrates a two-factor model with seven indicators. Each factor is associated with its respective observed items, with factor loadings for unrelated items set to zero. Additionally, we can determine factor correlations and error terms, which measure the residual variability. The two latent constructs could represent memory and cognitive function, with specific games serving as indicators for each construct based on their relevance to these functions. Games that primarily assess memory skills could be linked to the memory construct, while those evaluating cognitive processing, attention, or executive function could correspond to the cognitive function construct.

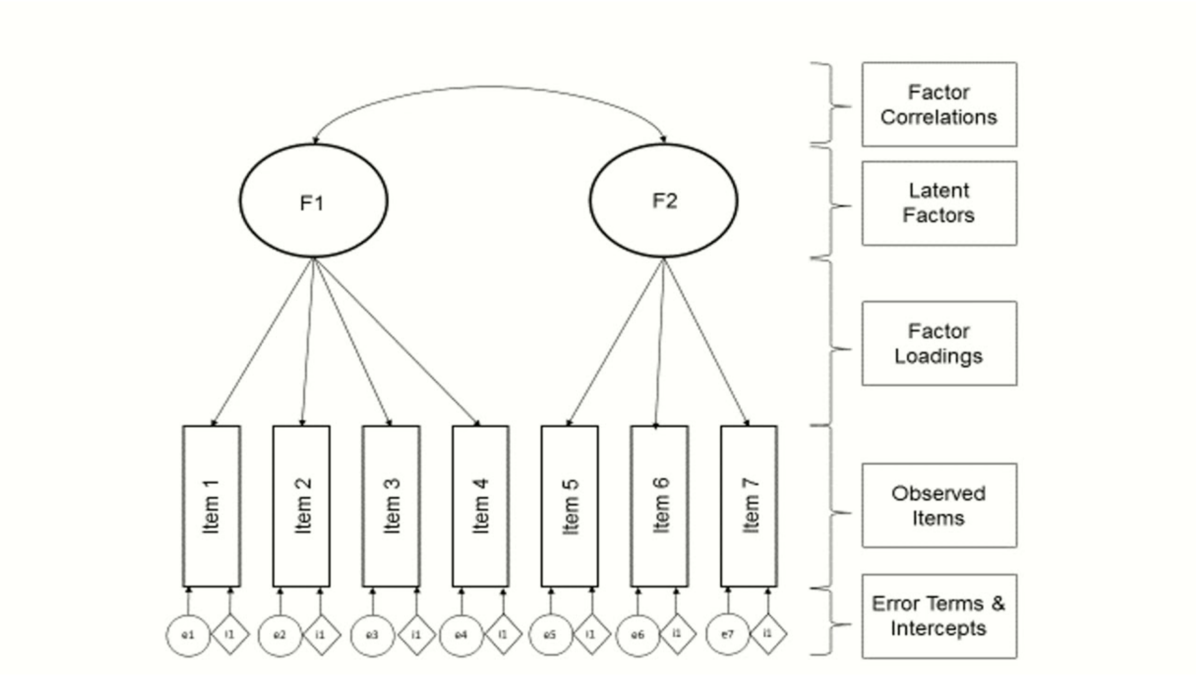


Figure 2.3: 2-factor model with 7 indicators.

Latent change score model

To implement latent change score modeling (LCSM), we use observations from cognitive games and brain volumes at two time points, spaced one year apart. In the previous section, we introduced a latent variable representing cognitive ability, measured using game scores as indicators. Next, we build univariate LCS models separately to the cognitive ability factor and the brain volume variable, giving the advantage to decompose observed scores into true scores and measurement errors. This approach provides a more accurate assessment of intra-individual change and inter-individual differences in both cognitive performance and brain structure. To construct a latent change score model with two time points, we express the latent variable at Time 2 as a function of the latent variable at Time 1, plus the latent change:[13]

$$L_{t2} = L_{t1} + \Delta L,$$

where L is the latent variable

Thus, to model latent change, we introduce a path from Time 1 to Time 2, fixing its factor loading to one, as indicated by the equation above. This model enables us to answer key research questions. Figure 2.4 illustrates the model structure and the relationships between observed and latent variables.

The most fundamental question is whether there is an average population-level change in cognitive ability and brain volumes between Time 1 and Time 2. The mean of the latent change scores provides the answer to this question, while the variance of the latent change scores indicates how much individuals differ in their changes over time. Additionally, we can include an autoregressive parameter to examine whether the amount of change depends on the initial score at Time 1.

Figure 2.4 presents the multiple-indicator latent change score model, where COG1 and COG2 represent latent cognitive ability at Time 1 and Time 2, respectively. The squares denote observed variables (scores from three different cognitive games), with factor loadings labeled as λ_1, λ_2, and λ_3. The latent change score (ΔCOG) is included in the model, with its own intercept and variance. To assume that time points are equidistant for all participants, the autoregressive path between COG1 and COG2 is fixed to unity. However, latent change score models (LCSMs) offer flexibility to accommodate non-equidistant time points, allowing for variability in time intervals across individuals.

Additionally, we account for task-specific consistency by allowing the error terms of each observed score to be correlated across time points. This is represented in the model by double-headed arrows connecting the same task at different time points. These correlations acknowledge that certain stable task-specific influences persist over time but are not necessarily part of the broader cognitive ability being measured. For example, a participant who has prior experience with tasks similar to the Stroop test may complete it more quickly at both Time 1 and Time 2, regardless of actual cognitive change. By incorporating these residual covariances, the model ensures that the analysis focuses on true cognitive changes rather than stable individual differences in task-specific performance.[13]

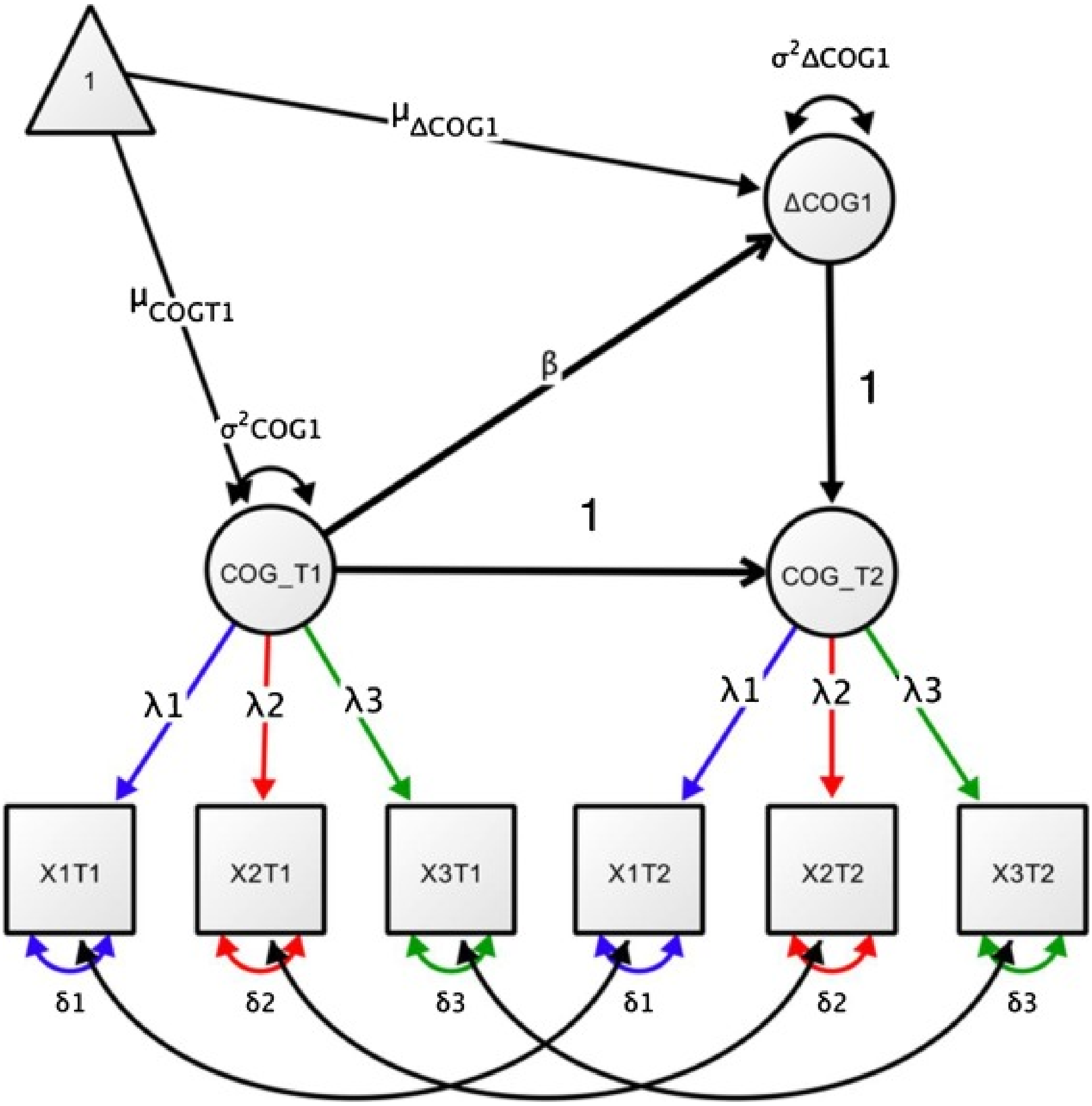


Figure 2.4: Multiple indicator Latent Change Score model.

One might argue that we could simply use the observed difference scores to build models based on them. However, this approach introduces significant measurement error. If we do not account for these errors, they will be present in both time points, ultimately reducing the reliability of the change score. To address this issue, we use the observed brain volume measurements from Time 1 and Time 2 to define a latent change score. This approach is essentially equivalent to a paired t-test but provides a more robust framework by explicitly modeling the underlying change while accounting for measurement error. Figure 2.5 illustrates the structure of the latent change score model for brain volume.

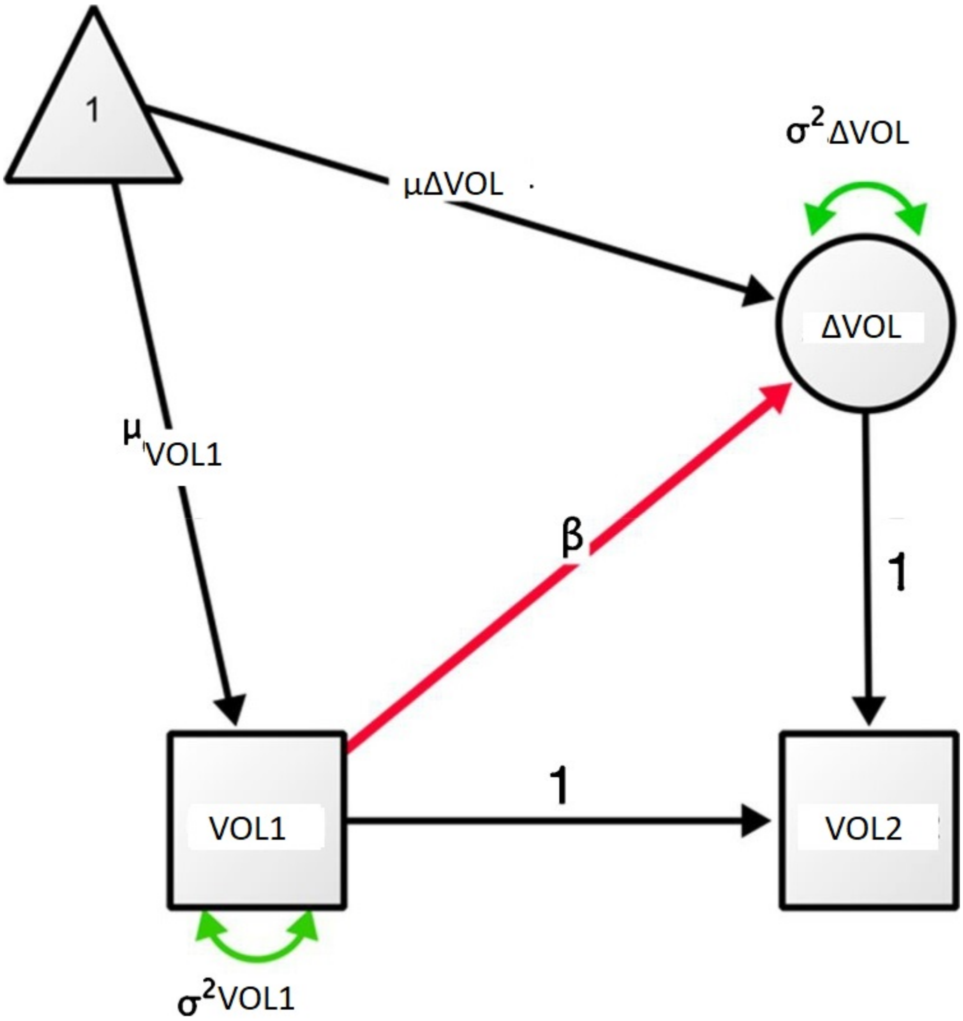


Figure 2.5: Latent Change Score model for brain volume.

Bivariate Latent Change Score Model

We have introduced the key components of the Bivariate Latent Change Score Model. Now, we can connect these pieces and derive some insightful conclusions. Figure 2.6 illustrates the complete model that we will use to examine the relationships between game performance and brain volumes. As we can see, this model captures four important aspects of brain volume–game performance interactions. First, it includes the covariance between them at baseline, which informs us about their linear relationship. For example, if this metric is positive, individuals with higher game scores will tend to have larger brain region volumes. Second, we investigate how cognitive performance and brain volume are linked over time, specifically whether game performance at Time 1 influences the rate of change in brain structure. Third, we explore whether initial brain volume levels predict the degree of change in game performance. Lastly, we assess correlated changes, examining the extent to which alterations in brain volume and game performance occur together after accounting for their directional influences.[13]

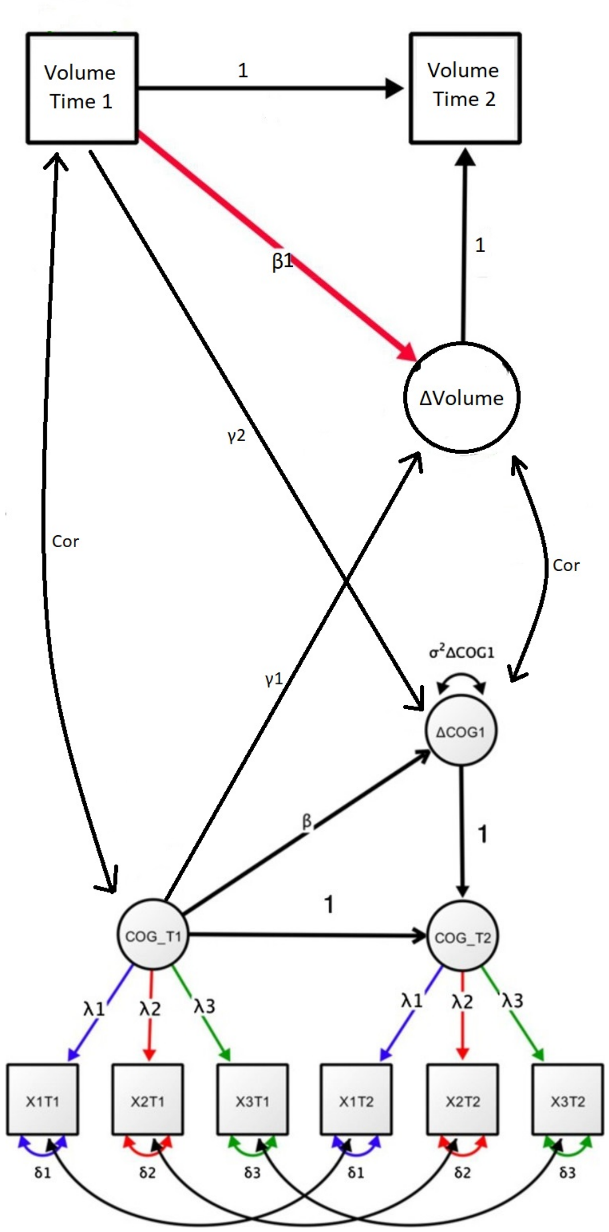


Figure 2.6: Bivariate Latent Change score model for brain volume and games' performance.

Estimates and Model fit

After the decision on the model we will use and the hypothesis that we want to test, the estimation of the parameters follows. The method that is used in Structural Equations Modeling is the Maximum Likelihood (ML), which assumes that the data follow a multivariate normal distribution. When we face real-world data, though, it is rarely possible that the data satisfies this assumption. Often, some variables may be skewed, have outliers or show heavy tails. That's why there have been developed other estimations that can handle this issue. Maximum Likelihood with Robust Standard Errors (MLR) adjusts for non-normality by computing robust standard errors. MLR does not assume perfect normality and the model fit assessments will be more reliable. After the estimation of the parameters we need to check how well the model fits the data. The way this works in SEM is by comparing the model-implied covariance matrix with the observed covariance matrix. So the covariance matrix that the model predicts and the covariance matrix that we actually see in the data. If these two matrices deviate a lot, then it indicates that the model might not be a good representation of reality. The most common model fit indices that have been used are presented in Table 2.2. The Chi-Square and Standardized Root Mean Square(SRMR) tests whether the model's predictions significantly differ from the observed data. However, it is highly sensitive when the sample size is big or we have many parameters. Comparative Fit Index and Tucker-Lewis Index compare the model to a baseline model. If the model fits the data significantly better that the model with no structure, then these indices will have a value bigger than 0.95. Lastly, Root Mean Square Error of Approximation (RMSEA) estimates how much error we expect per degree of freedom in the model. It is particularly useful for models with many parameters because it adjusts for model complexity and the validation of model fit is more accurate.[9] Likelihood Ratio Test (LRT): Used when comparing nested models (where one model is a simpler version of another). A significant chi-square difference suggests the more complex model fits better. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): These indices penalize overly complex models. Lower AIC/BIC values indicate a better balance between fit and simplicity.

Index	Good Model Fit Suggestion
Chi-Square Test	p-value >0.05
Comparative Fit Index	Values >0.95
Tucker-Lewis Index	Values >0.95
RMSEA	Values <0.06
SRMR	p-value <0.08

Table 2.2: Model Fit indices for SEM

Measurement Invariance

It is important to establish that the sample comes from a homogeneous population. Otherwise, the parameter estimates will not be valid for drawing conclusions. In reality, though, it is possible that the sample include participants from different subpopulations, such as different gender group. In Structural Equation Modeling (SEM), there is a straightforward way to compare parameter estimates across groups.

We impose equality constraints on the parameters of interest (e.g., factor loadings, intercepts). Then, we compare the constrained model (where parameters are assumed equal across groups) to the freely estimated model (where parameters are allowed to differ). If the constrained model shows a significantly worse fit, this suggests that the parameters differ across groups, indicating meaningful statistical differences. This method is particularly useful for improving the reliability of parameter estimates and testing specific hypotheses. A key step in longitudinal SEM and multigroup comparisons is establishing measurement invariance—ensuring that the same latent construct is measured consistently across time or across groups. Without measurement invariance, comparing latent constructs across groups is meaningless, as differences may reflect measurement biases rather than true differences.

Why is this important? For example, there could be some bias in the brain volumes of males population in comparison to females, that could obscure the true differences in groups. This issue, known as differential item functioning (DIF), introduces testing bias, meaning that observed differences may be due to the measurement process rather than actual cognitive ability. Thus, before comparing groups, researchers must first establish measurement invariance to ensure that the latent construct is being measured in the same way across populations. We have different types of measurement invariances that we want to compare and derive conclusions from them:

Configural Invariance: The same factor structure holds across groups.

Weak Invariance: Factor loadings are equal across groups, meaning that the relationships between indicators and the latent construct are the same.

Scalar Invariance: Factor loadings and intercepts are equal, allowing for valid mean comparisons of the latent construct across groups.

Strict Invariance: Factor loadings, intercepts, and residual variances are equal across groups, ensuring that measurement error is also consistent. While weak invariance confirms that the same construct is being measured, it does not allow for latent mean comparisons. Scalar invariance is necessary for valid comparisons of latent constructs across groups or over time.

In a longitudinal study, we want the relationship between latent variables and observed scores to remain constant across time, even if the latent scores themselves increase or decrease due to natural cognitive change. Failing to establish measurement invariance can lead to misleading conclusions about cognitive change, brain volume trajectories, and their associations with other variables. Before making any group comparisons or longitudinal conclusions, it is essential to establish at least scalar invariance. Without it, observed differences in cognitive ability, brain volume, or other constructs may reflect inconsistencies in measurement rather than real changes or group differences.[15]

Missing Values

There is a way to handle missing values efficiently in Structural Equation Modeling. When data are Missing Completely At Random (MCAR) or Missing At Random (MAR), meaning that missingness is related only to observed variables in the dataset , Full Information Maximum Likelihood (FIML) can be used to estimate model parameters using all available data. The main benefits of FIML, in comparison to listwise deletion and other methods, is the maximization of the statistical power. Instead of deleting cases with missing values, FIML retains all the available data and incorporate them into the likelihood function. It has been proven that FIML performs the same of better than Multiple Imputation, another method that resolves the missing values problem in MAR occasions. A practical advantage of FIML is that it produces stable estimates across repeated analyses, whereas MI relies on stochastic sampling, meaning that results may vary slightly across different imputations. However, MI has its own strengths, particularly when the missingness mechanism is complex or involves variables that are not included in the model, in which case imputation models can incorporate additional external information.[13]

Summary

In this chapter, we explored the theory behind Mixed-Effects Models and Structural Equation Modeling (SEM). Both are well-suited for longitudinal data, as they provide the necessary tools to incorporate repeated measurements into the analysis. These models are particularly valuable for dementia research, as they allow us to examine both intra-individual and inter-individual differences, helping to identify factors that influence the disease. In the next chapter, we will apply these models to investigate the relationship between cognitive game performance and brain volume changes, drawing meaningful inferences from the data.

Chapter 3

Descriptive Analysis

3.1 Experimental Structure

Researchers from the University of California, San Francisco, along with ALLFTD team have developed a mobile application to gather additional data from patients with frontotemporal dementia. This neurodegenerative disease affects many individuals, and, as of now, there is no cure. Therefore, further research and increased data collection from patients are crucial. Traditionally, diagnostic tests like brain MRI scans and biofluid analysis have been the primary sources for researchers seeking potential treatments and evaluating the severity disease. However, these methods are both challenging to conduct and costly. To address these challenges, researchers at UCSF and ALLFTD created an application that includes games designed to assess cognitive and memory functions.[20] This app features five different games, and our dataset includes both the games' performance and the results from eight brain volumes from MRI scans.

Participants

Data from 279 participants are gathered from an ongoing research project on Frontotemporeal dementia that are conducted by multiple centers around the world. It contains healthy controls, asymptomatic f-FTD mutation carriers and symptomatic patients. More specifically, there are FTD patients and asymptomatic individuals from f-FTD families with a known C9orf72 expansion or pathogenic mutation in GRN or MAPT. In addition, mutation carriers and non-carrier family controls were included. People that have these genes tend to have less brain volumes than those with healthy minds. Therefore, it is reasonable to have some issues in cognitive and memory tasks.

Study Design

Participants play each game three times over a 12-day interval every 6 months, with washout days between sessions. Sessions were executed remotely via a smartphone app. The reason to test the games multiple times is to assess the reliability of the games as a measurement.[2] During this period, they underwent MRI scans in order to obtain the brain volumes and some other features of the brain. However, it seems that even though games were played every 6 months, our data contains volumes over a year. To ensure the validity of the analysis, it is necessary to adjust the data by using only the game scores most representative of the brain volumes recorded. Specifically, we will focus on the game scores collected closest to the time of the MRI scans, limiting the analysis to data from the same year. In addition, it is important to mention that we will use only the first time that participants played the game. This is because volumes do not change in 12-day interval and thus there is not useful information if we keep all of them in order to assess the association between games and volumes. Although calculating a statistical measure (e.g. an average) from the three scores could be an option, it may introduce bias, as some participants did not complete all three sessions. In this analysis, brain volumes will be used in comparison with the serious games and some demographic variables.

3.2 Smartphone Cognitive Testing

The serious games assess aspects such as memory, attention, and cognitive function. They are sensitive to these issues, even in patients in the prodromal stages. Let's review them one by one and describe how they work (Figure 3.1 and 3.2):

Stroop: A color word is displayed in ink of a different color, and the participant's task is to name the color of the ink, not the word itself. For example, if the word reads ”blue” but is printed in red ink, the correct response would be ”red.” Healthy individuals tend to have greater cognitive control, allowing them to identify the mismatch between the word and the ink color more quickly. The Stroop test is widely used to assess cognitive function in patients with frontotemporal dementia. After the participant completes the Stroop task, we collect data on the number of correct answers and the reaction times for each trial. From the median reaction times across all trials, we derive a variable that indicates a speed-accuracy score on a scale of 0 to 10 (more is better).

Flanker: Five ducks are displayed on the screen, with one main duck in the center surrounded by the others. The participant's task is to indicate the direction in which the main duck is facing. This can be challenging because the surrounding ducks may face the opposite direction, requiring increased focus to inhibit an automatic response. Similar to the Stroop task, we collect reaction time and accuracy data from all trials. We will focus to the speed accuracy score using total reaction time for all trials on a 0-10 range (more is better). This measurement combines reaction time and accuracy into a single score.

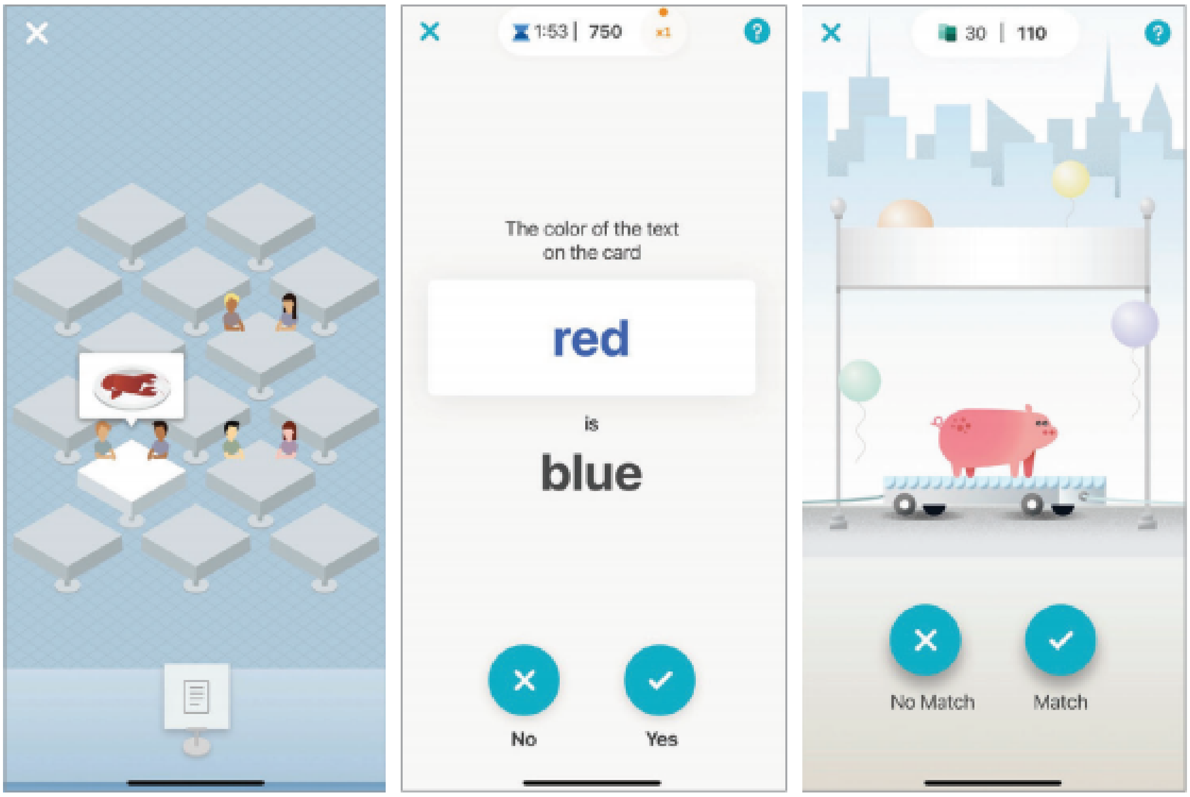
Humi: To keep all your new customers happy, you need to remember their orders and deliver the food to the correct tables. In this game, you must recall which dishes were ordered and which table placed each order. Patients with frontotemporal dementia often struggle with working memory and attention, making it challenging for them to remember this information. Our data includes the mean of correct trials, calculated by averaging the correct responses across all rounds.

Nback: You see animal floats go by one at a time. The goal is to indicate correctly if the animal that is presented 2 steps earlier in the sequence is the same. Like the Humi game, people with poor working memory and concentration would score worse than those with healthier minds. The data contains:

1. the total trials that participants match for a not match trial (false positive)
2. total match trials where participants gives a respond
3. total not match trials where participant gives a respond
4. total trials that participants respond match to a match trial (True positive)
5. hit rate= true positive responds/ total match trials
6. false rate= false positive responds/ total not match trials
7. Z-score hit rate - Z-score false rate

We will involve in our analysis the last variable, which represents perfectly the performance of the participants.

Card Shuffle: Cards are presented to the participant, each featuring shapes that may differ in color or number. In each round, the participant must match the card presented to them with one of four other cards. The correct match depends on a specific rule that the participant must remember in order to answer correctly (e.g., matching cards with the same shape). After several rounds, the rule changes, and the participant must adapt to the new rule until it changes again. This test assesses the ability to shift thinking in response to changing rules. The total number of correct trials will be used in our analysis.



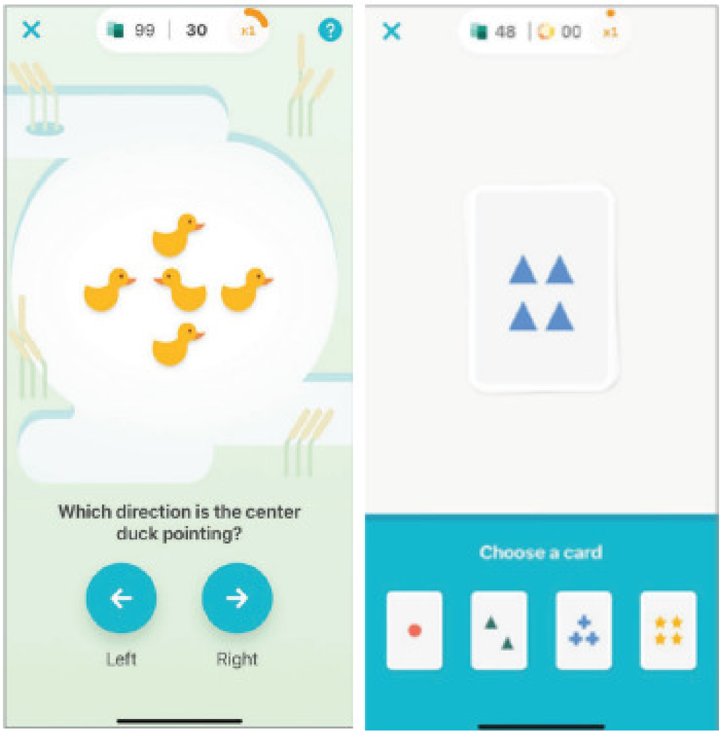
The color of the text on the card

red

is

blue

Figure 3.1: Top left:Humi, Top right:Nback, Top:Stroop



Which direction is the center duck pointing?

Choose a card

Figure 3.2: Down left:Flanker, Down right:Card shuffle

3.3 Brain Region Volumes from MRI

To validate whether the games described above can be effective for research on frontotemporal dementia, we first need to identify another metric that has been used for this purpose. Brain volume measurements from MRI imaging are a well-known method for assessing the presence of this disease. These volumes represent the sizes of specific brain regions, measured in cubic millimeters, and can be influenced by factors such as age, diseases, and other neurological conditions. These measurements are obtained through MRI scans with the help of specialized software that segments the brain into different regions and counts the number of 3D pixels (voxels).

In patients with frontotemporal dementia, certain brain regions may shrink as the disease progresses. Specifically, the areas most affected are the frontal and temporal lobes. The former is primarily responsible for behavior, personality, and decision-making, while the latter is more involved in language, memory, and emotions.[5] Therefore, for games that assess executive functions, we expect the frontal lobes to play the biggest role, while for games involving memory, the temporal lobes are likely to be the most important.

In more advanced stages of the disease, the parietal and occipital lobes may also be affected. These lobes are responsible for spatial awareness, sensory input, and visual processing. Our data includes brain volume measurements for these four lobes, further divided into left and right hemispheres. The left and right hemispheres serve distinct functions, and the disease may have different effects depending on where the atrophy occurs. The left hemisphere is more involved in language, logic, and analytical thinking, while the right hemisphere is associated with emotions, spatial awareness, and intuition. We anticipate that the games will be more correlated with the volumes of the left lobes than the right.

Finally, we normalized the brain volume data by dividing the measurements by the total intracranial volume to account for differences in overall brain size among individuals. Thus, our response variables are the normalized volumes of the four lobes (left and right), measured in cubic millimeters. We expect these measurements to decrease over time due to either the progression of the disease or natural aging.

Severity Disease Measurement

Other than brain volumes data, there is a measurement that is mostly used to evaluate the stage of the disease in patients. Basically, it is an aggregation of data coming from interviews with patient and caregiver combined with biomarkers and MRI scans.

This measurement is based on the clinical dementia rating scale (CDR) and the National Alzheimer's Coordinating Center(NACC). CDR is a tool to evaluate cognitive and functional abilities such as memory, orientation, problem solving and community affairs. Then, they assign a score in each domain ranging from 0 to 3.

- **0:** Asymptomatic or Healthy
- **0.5:** Prodromal stage
- **1-3:** Symptomatic

The NACC FTLD Module complements the CDR by testing features that are more specific for Frontotemporal Dementia like behavioral changes, language deficits and motor symptoms. This metric is widely used for diagnosing and monitoring progression and for research studies. [20]

3.4 Exploratory Data Analysis

In this study, brain volumes are used to evaluate the validity of cognitive games as a digital biomarker for frontotemporal dementia. Any information extracted from these mobile games that can help assess the stage of the disease will be invaluable for clinicians. Magnetic resonance imaging (MRI) offers several advantages over other dementia assessment methods. Unlike the Clinical Dementia Rating (CDR) and the National Alzheimer's Coordinating Center (NACC) measurements, which rely on subjective reports from caregivers and clinicians, MRI provides objective data. Additionally, tracking brain volume changes over time allows clinicians to monitor the progression of dementia-related deterioration, making it a valuable tool for both research and treatment evaluation. In this analysis, we use brain volume measurements to examine the overall association between game performance and brain volumes, as well as how these relationships evolve over time.

Brain Volumes as Biomarker for Frontotemporal Dementia

Brain volumes are widely known and used by clinicians and the changes over time are a valuable asset for them. We can test their validity as biomarker for frontotemporal dementia by observing whether they can differentiate participants between the stages from CDR and NACC measurement. In Figure 3.3 we can see the number of participants in each severity stage. We would like to see whether there are differences between the brain volumes of each group in CDR and NACC measurement. Figure 3.4 shows the boxplots of left frontal lobe by severity stage. From boxplots and some statistical tests (Kruskal-Wallis $p-value < 0.01$ and ANOVA with unequal variances $p-value < 0.01$), we can assume that there are differences in the brain volume between patients with different severity stage. By implementing pairwise comparisons using Wilcoxon rank sum test, it seems that every pair of stages are statistically different. The same results apply for the other brain volumes.

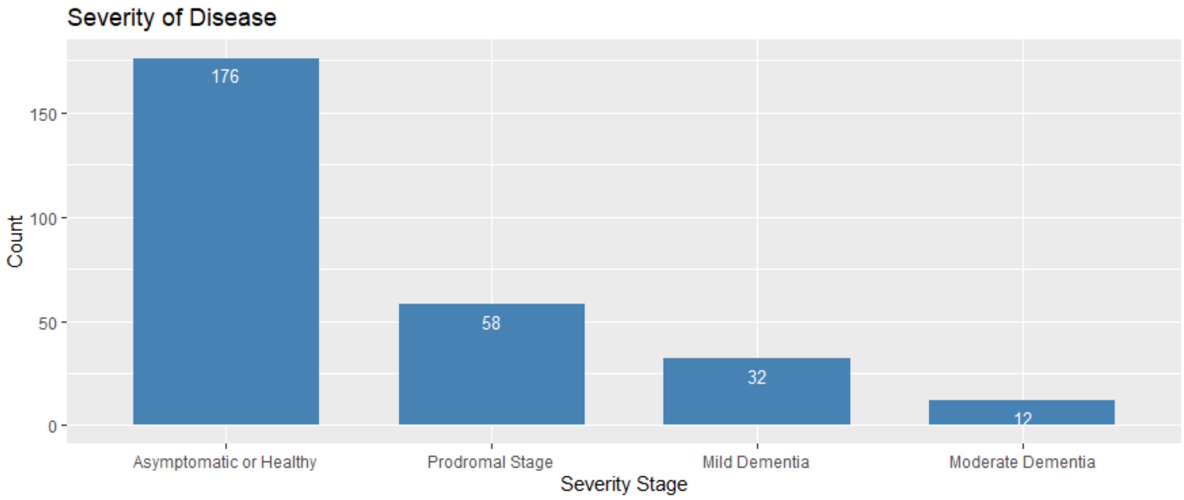


Figure 3.3: Number of participants in each stage of CDR and NACC measurements

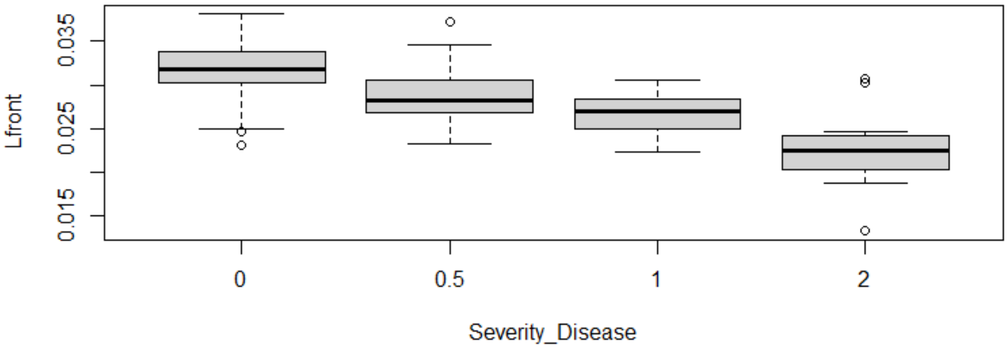
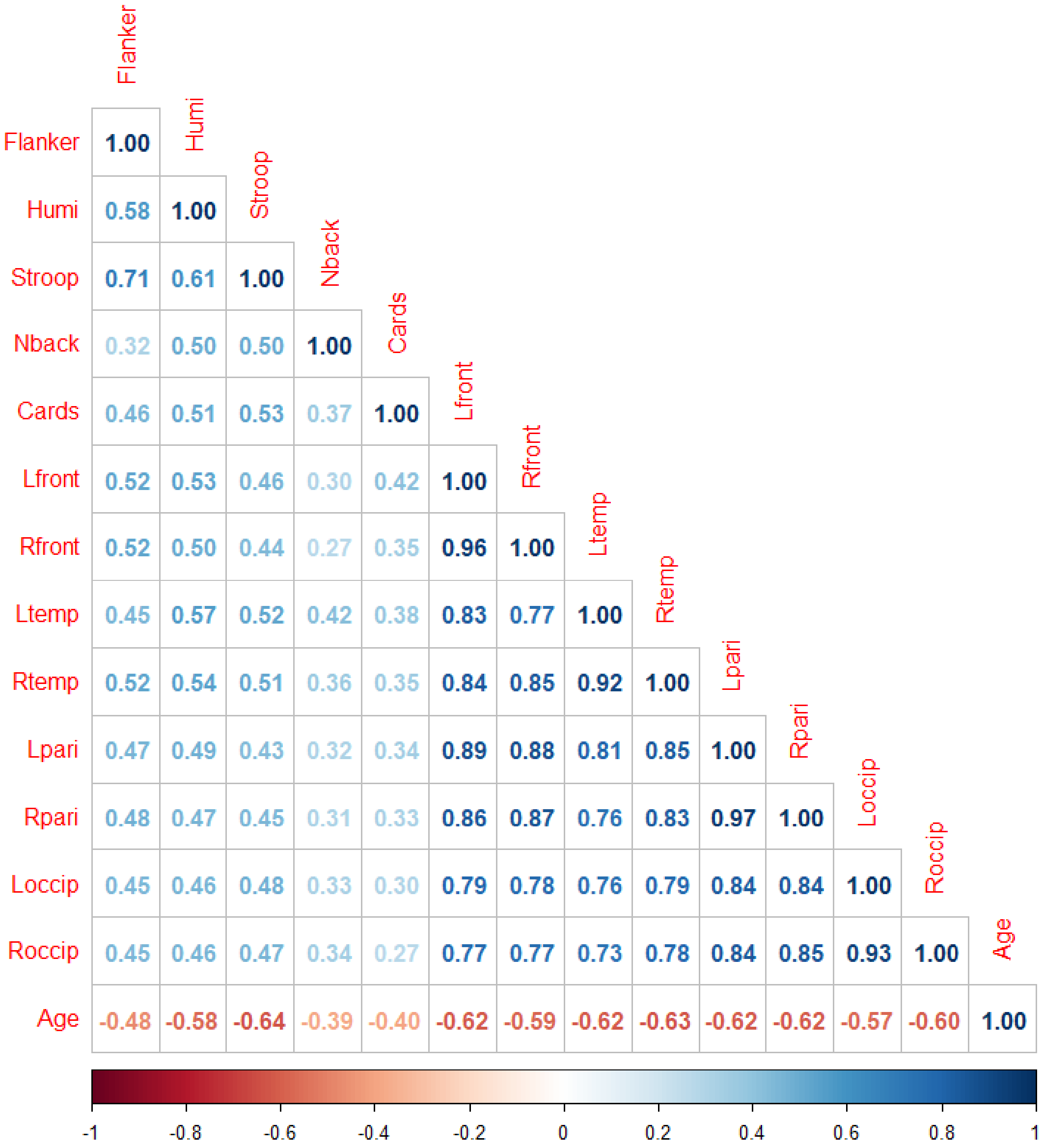


Figure 3.4: Differences in volumes by Severity Stage

Figure 3.5 and Table 3.1 provide basic information about the variables and their correlations. The mobile games show a mild positive correlation with brain volumes, ranging from 0.3 to 0.6. Among the games, Flanker, Humi, and Stroop exhibit the strongest correlations, while NBack and Cards Shuffle show weaker correlations. Age appears to be negatively correlated with both game performance(between -0.4 and -0.65) and brain volumes(between -0.57 to -0.63). These findings suggest that participants with higher game scores tend to have larger brain volumes, whereas older participants have smaller brain volumes compared to younger ones. Furthermore, there are strong correlations observed between brain volumes, particularly between the same left and right lobes. From Table 3.1, we can observe that participants are between ages of 18 and 85. Moreover, it is important to mention that the lobes' volumes are normalized by the total intracranial volume, in order to account for individual differences of brain sizes. Practically, this means that the brain volumes in Table 3.1 are the proportions of the regions over the total intracranial volume. Finally, the right lobes have about the same values like the left lobes, so they are not included in the Table 3.1 below.



	Flanker	Humi	Stroop	Nback	Cards	Lfront	Rfront	Ltemp	Rtemp	Lpari	Rpari	Loccip	Roccip	Age
Flanker	1.00													
Humi	0.58	1.00												
Stroop	0.71	0.61	1.00											
Nback	0.32	0.50	0.50	1.00										
Cards	0.46	0.51	0.53	0.37	1.00									
Lfront	0.52	0.53	0.46	0.30	0.42	1.00								
Rfront	0.52	0.50	0.44	0.27	0.35	0.96	1.00							
Ltemp	0.45	0.57	0.52	0.42	0.38	0.83	0.77	1.00						
Rtemp	0.52	0.54	0.51	0.36	0.35	0.84	0.85	0.92	1.00					
Lpari	0.47	0.49	0.43	0.32	0.34	0.89	0.88	0.81	0.85	1.00				
Rpari	0.48	0.47	0.45	0.31	0.33	0.86	0.87	0.76	0.83	0.97	1.00			
Loccip	0.45	0.46	0.48	0.33	0.30	0.79	0.78	0.76	0.79	0.84	0.84	1.00		
Roccip	0.45	0.46	0.47	0.34	0.27	0.77	0.77	0.73	0.78	0.84	0.85	0.93	1.00	
Age	-0.48	-0.58	-0.64	-0.39	-0.40	-0.62	-0.59	-0.62	-0.63	-0.62	-0.62	-0.57	-0.60	1.00

Figure 3.5: Correlation between variables

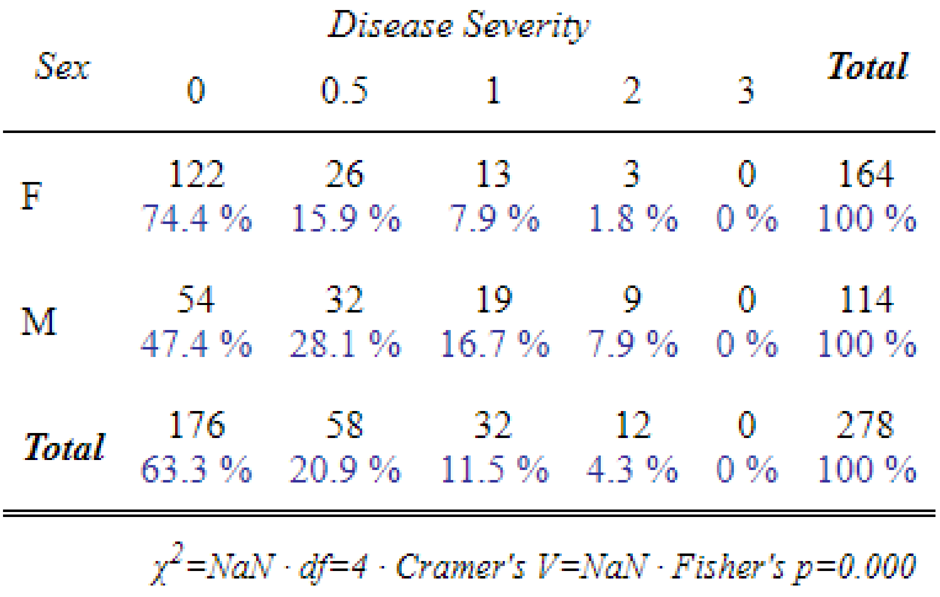
	Mean	St. Deviation	Median	Min	Max
Flanker	**8.28**	**1.22**	**8.53**	**1.01**	**9.78**
Humi	**3.00**	**1.01**	**3.00**	**0.20**	**5.78**
Stroop	**6.87**	**1.10**	**7.05**	**3.51**	**9.12**
NBack	**1.96**	**1.11**	**2.04**	**-0.36**	**6.18**
Card	**33.28**	**7.20**	**36.00**	**8.00**	**42.00**
Age	**51.60**	**14.45**	**51**	**18**	**85**
Left Frontal	**0.030**	**0.004**	**0.031**	**0.013**	**0.038**
Left Temporal	**0.021**	**0.003**	**0.022**	**0.009**	**0.027**
Left Parietal	**0.018**	**0.002**	**0.018**	**0.008**	**0.023**
Left Occipital	**0.007**	**0.001**	**0.007**	**0.003**	**0.009**

Table 3.1: Statistics

Gender-Volume Association

Out of the 278 participants, 164 are female and 114 are male. After conducting statistical tests, we rejected the hypothesis that brain volumes are the same across genders (two-sample t-test, $p-value < 0.01$). However, there is an important consideration. As shown in Figure 3.6, the distribution of men and women across severity stage groups reveals that men with CDR and NACC scores above 0.5 outnumber women in percentage terms.

When adjusting for those who are healthy, a gender difference remains, but it is not as pronounced, as illustrated in Figure 3.7 (two-sample t-test, $p - value < 0.01$).



Sex	*Disease Severity* 0	0.5	1	2	3	***Total***
F	122 74.4 %	26 15.9 %	13 7.9 %	3 1.8 %	0 0 %	164 100 %
M	54 47.4 %	32 28.1 %	19 16.7 %	9 7.9 %	0 0 %	114 100 %
Total	176 63.3 %	58 20.9 %	32 11.5 %	12 4.3 %	0 0 %	278 100 %

χ^2=NaN · df=4 · Cramer's V=NaN · Fisher's p=0.000

Figure 3.6: Amount of males and females by severity disease

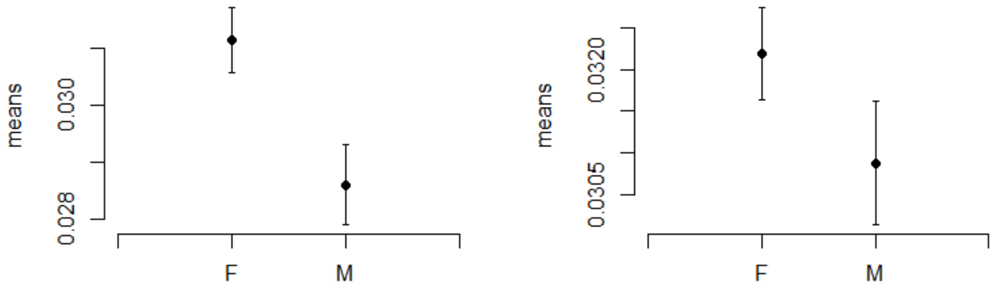


Figure 3.7: Gender-based differences in volumes. Left one for all the participants and the right one for only the healthy participants

Age-Brain Volume Association

As mentioned before, it is natural for brains to atrophy as we age, so it is expected that brain volumes decrease over time. In addition, dementia typically develops in older ages, which means that older people are more likely to be diagnosed with the disease. As we can see in Figure 3.8, there is a decreasing relationship between age and left frontal lobe volume. Moreover, almost all individuals who have symptoms are after the age of 40 with some exceptions that are probably due to some genetic factors.

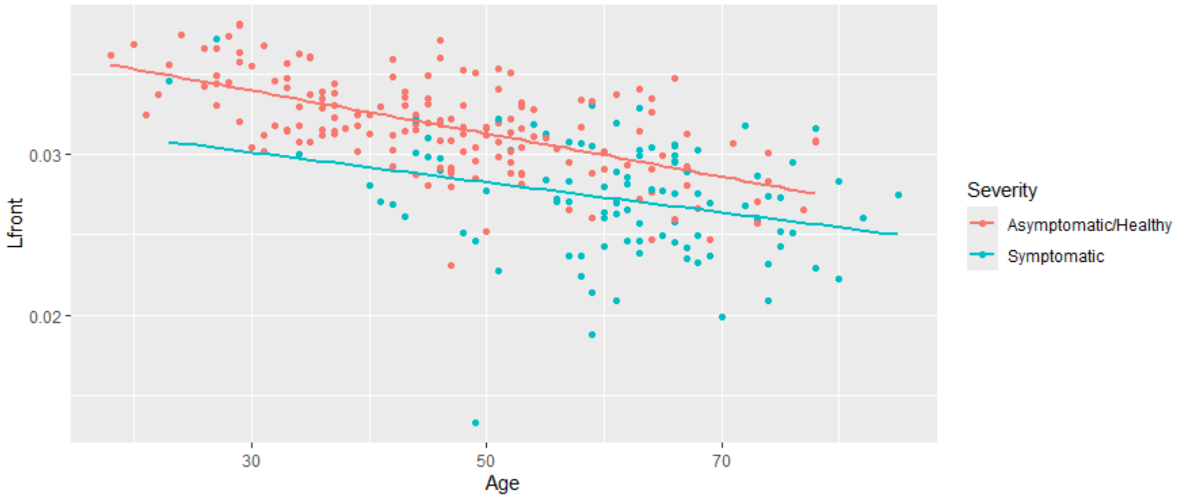


Figure 3.8: Age-Brain Volume relationship

Cognitive Games-Brain Volume Association

The primary goal of this study is to explore the relationship between game performance and brain volumes. These games are designed to assess participants' memory and cognitive function. It would be reasonable to assume that individuals who perform better in these games have healthier brains, which would likely correlate with larger brain volumes. Figures 3.9, 3.10 and 3.11 illustrate the relationship between game performance and left frontal lobe volume, distinguishing between healthy/asymptomatic and symptomatic participants. In the Flanker game, symptomatic participants consistently score lower than their asymptomatic counterparts. This trend is also evident in other games.

Figures 3.12–3.16 present scatter plots for each game and left frontal lobe volume, categorizing participants by three age groups and health status. In the Flanker game, healthy individuals maintain high scores regardless of age. In contrast, symptomatic participants show greater variability, with younger individuals generally scoring higher than older ones. In the Humi game, symptomatic patients consistently score lower than healthy participants. Moreover, the relationship between game performance and brain volume is more clearly observed among symptomatic individuals, suggesting a stronger link between cognitive decline and brain atrophy. In the Stroop game, older adults tend to perform worse across both healthy and symptomatic groups. Additionally, symptomatic participants score lower than healthy ones, similar to the trend seen in the Humi game. Finally, in the N-back and Card games, the relationship between game performance and brain volume is less distinct. However, symptomatic participants still score lower than healthy individuals, reinforcing the overall trend of cognitive decline being associated with lower brain volumes.

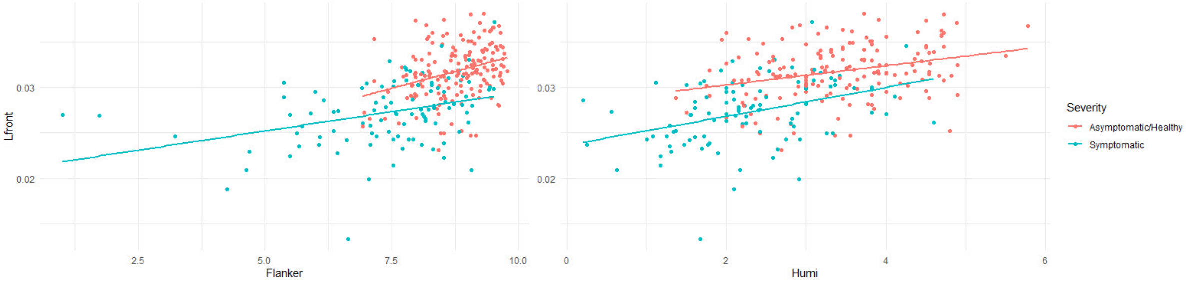


Figure 3.9: Games-Volume relationship

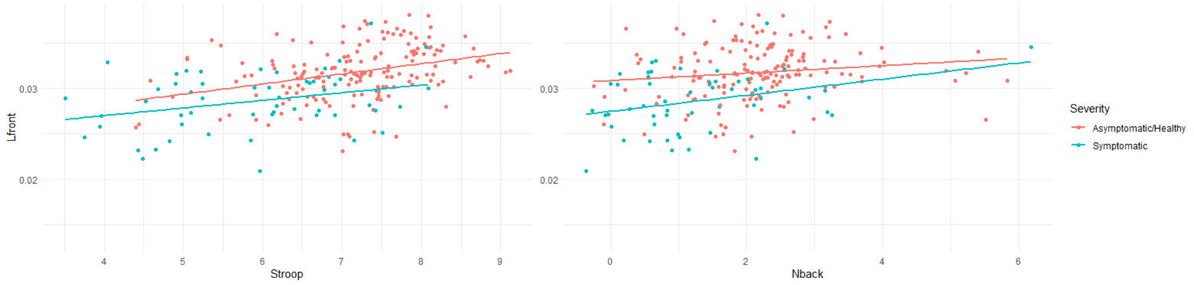


Figure 3.10: Games-Volume relationship

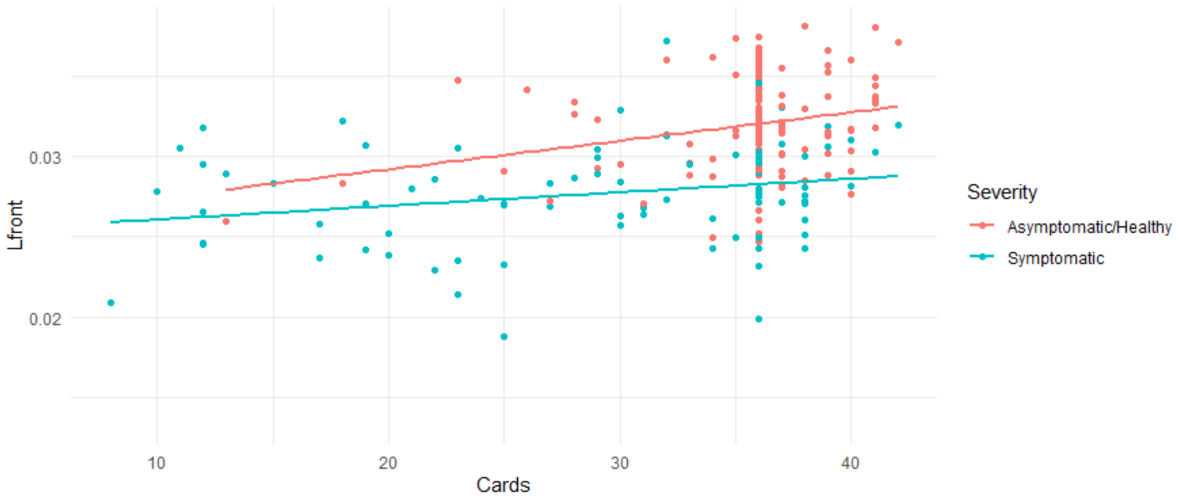


Figure 3.11: Cards-Volume relationship

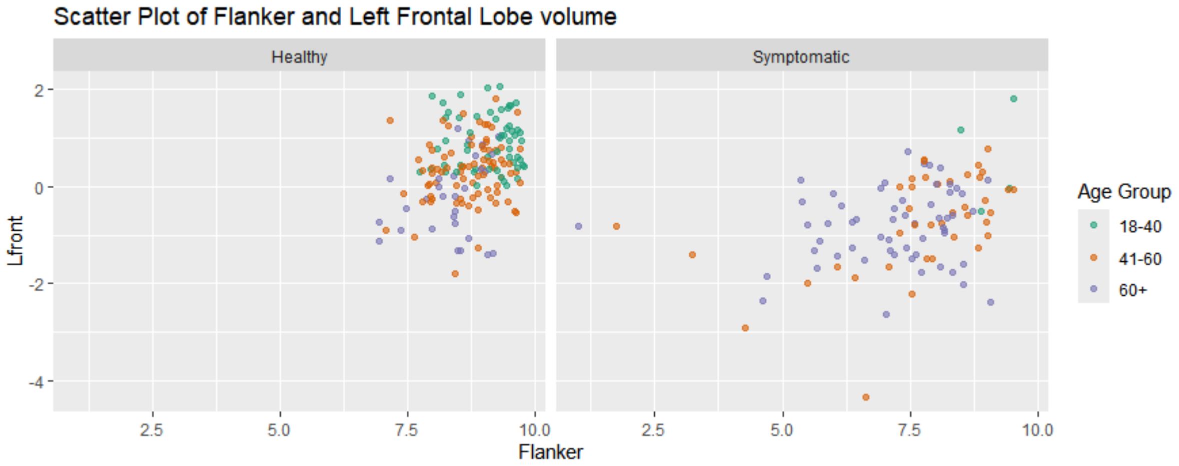


Figure 3.12: Relationship between Flanker and Left Frontal Lobe volume by age and health statement

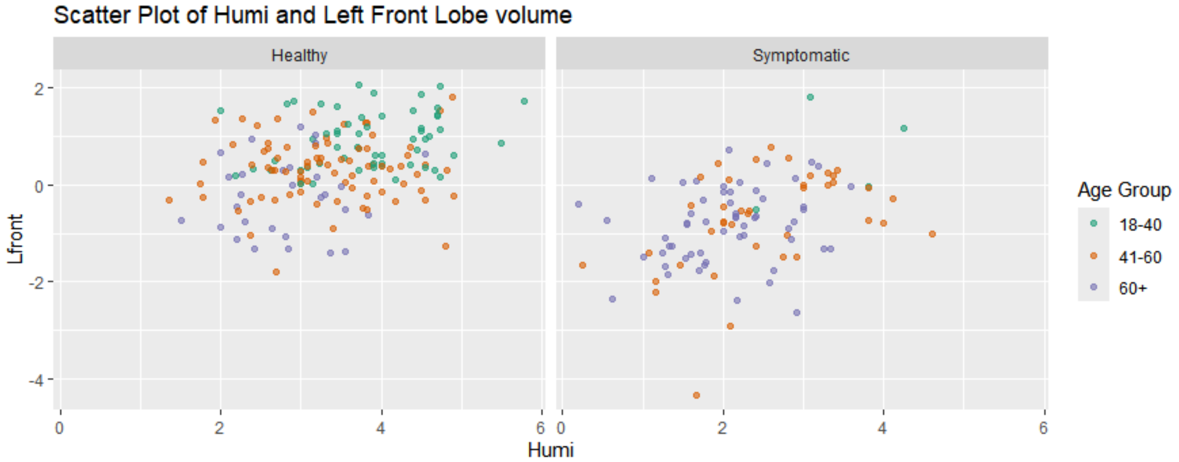


Figure 3.13: Relationship between Humi and Left Frontal Lobe volume by age and health statement

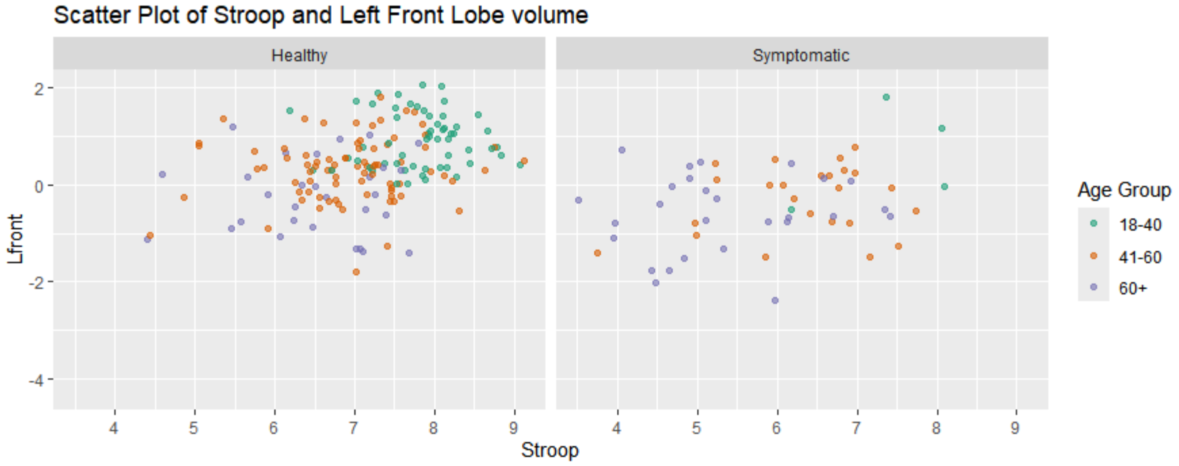


Figure 3.14: Relationship between Stroop and Left Frontal Lobe volume by age and health statement

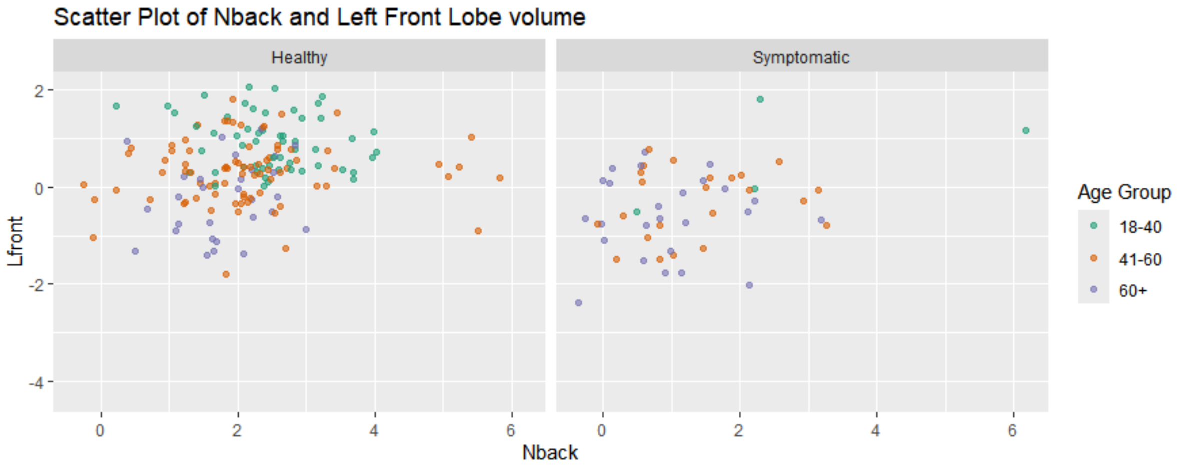


Figure 3.15: Relationship between Nback and Left Frontal Lobe volume by age and health statement

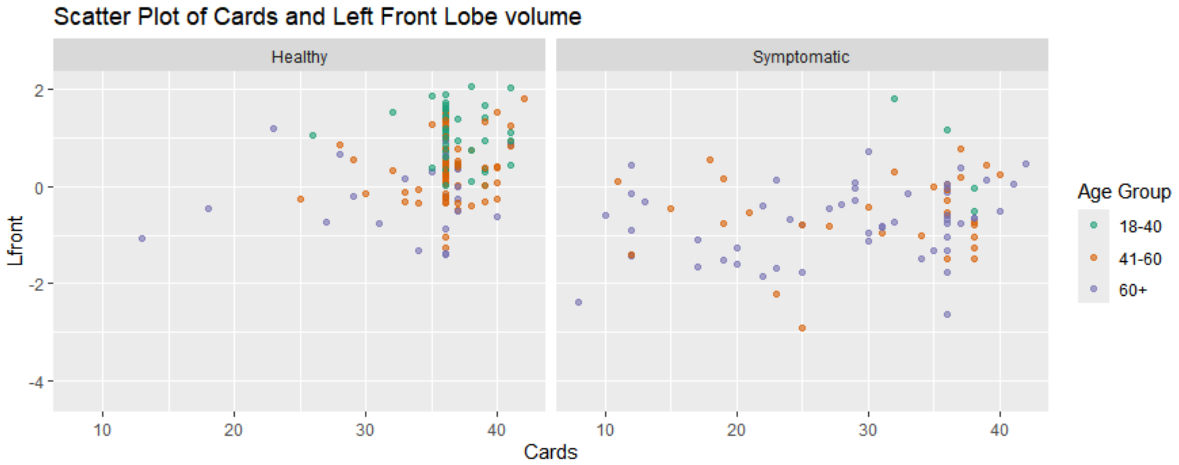


Figure 3.16: Relationship between Cards and Left Frontal Lobe volume by age and health statement

Brain Volumes over time

The dataset includes 206 participants with one time point, 71 with two time points, and 2 with three time points. Table 3.2 displays the severity stages of the 73 participants with more than one time point. After performing a Wilcoxon signed-rank test on the differences between time 1 and time 2 for the 73 participants, we found that we cannot reject the null hypothesis of no significant difference between the two time points ($p-value > 0.01$).

Figure 3.17 displays the histogram of percentage changes in left frontal lobe volume over time. As shown, most participants exhibit small changes, either positive or negative, with a few exceptions. Specifically, the average percentage change in left frontal lobe volume is -1.35%. The minimum and maximum percentage changes are -13.3% and +9.31%, respectively. Other brain volumes exhibit similar trends, as shown in Figure 3.18.

	Severity Stage			
	Healthy	**Prodromal**	**Mild Dementia**	**Moderate Dementia**
# of participants with two time points	44	19	6	4

Table 3.2: Participants with two different time points

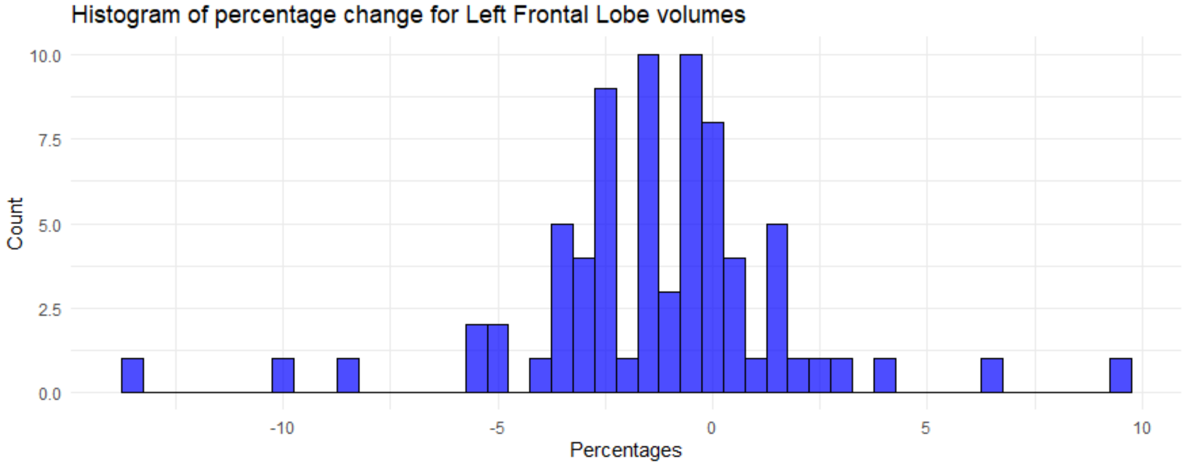


Figure 3.17: Percentage of changes over a year for Left Frontal lobe volume

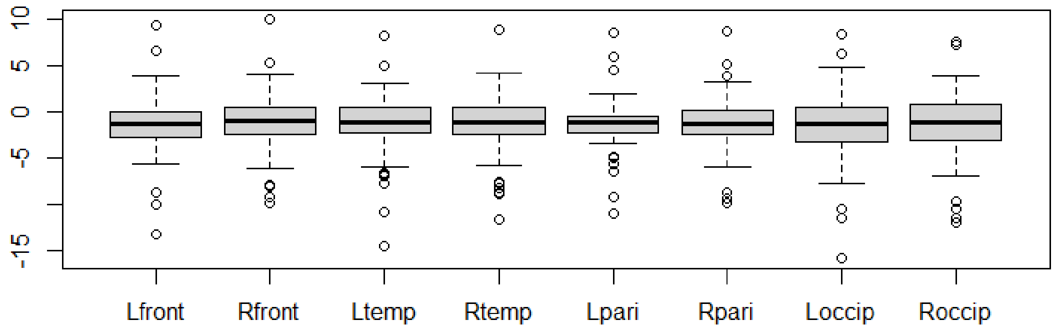


Figure 3.18: Boxplots indicating the distribution of percentage changes in volumes over a year for every brain volume.

Age-Brain Volume Association over time by Health Statement

Figure 3.19 illustrates the relationship between age and the percentage of change in left frontal lobe volume between two time points. The data suggest that the decline in symptomatic participants is slightly greater than in healthy individuals. However, since symptomatic participants are older, we can not know for sure whether this greater decline is due to their health condition or simply because of age. Therefore, we cannot draw definitive conclusions about the underlying cause of the increased rate of decline.

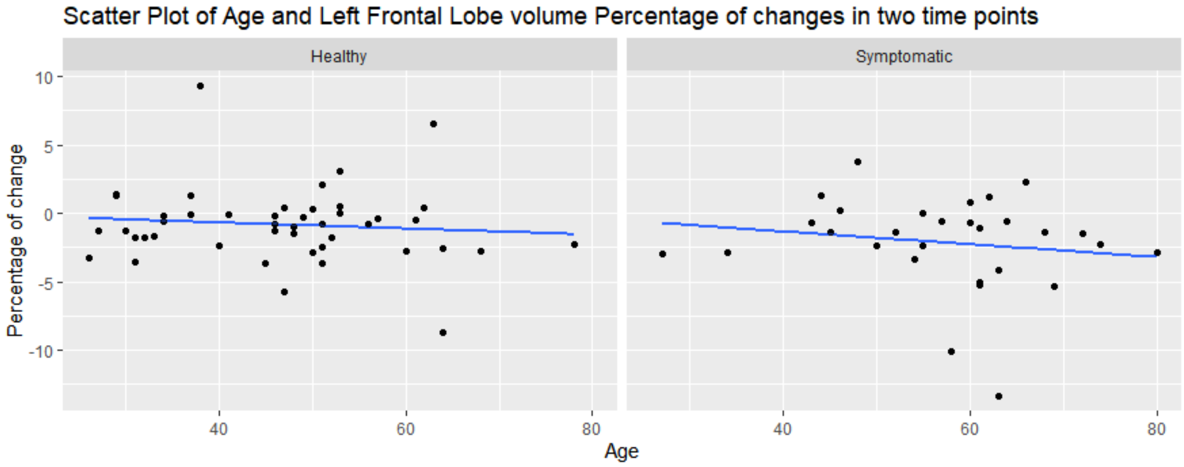


Figure 3.19: Association between Age and Percentage changes of Left Frontal Lobe Volume by health statement.

Cognitive Games over time

In contrast to brain volume percentage changes, cognitive game performance exhibits greater variability. Since several factors can influence an individual's performance—such as mood, familiarity with the games, and distractions—the observed fluctuations are reasonable. Brain volume measurements, on the other hand, have significantly lower measurement error compared to cognitive games, which is why such variations are not as evident. Furthermore, as shown in Table 3.3, the mean and median generally indicate a positive change over time, except for the Flanker and Cards games, which contrast with the brain volume results. This discrepancy may be due to the higher proportion of healthy individuals in the dataset. If healthy participants tend to improve their performance over time, which would be expected, it could lead to an overall positive percentage change in game performance.

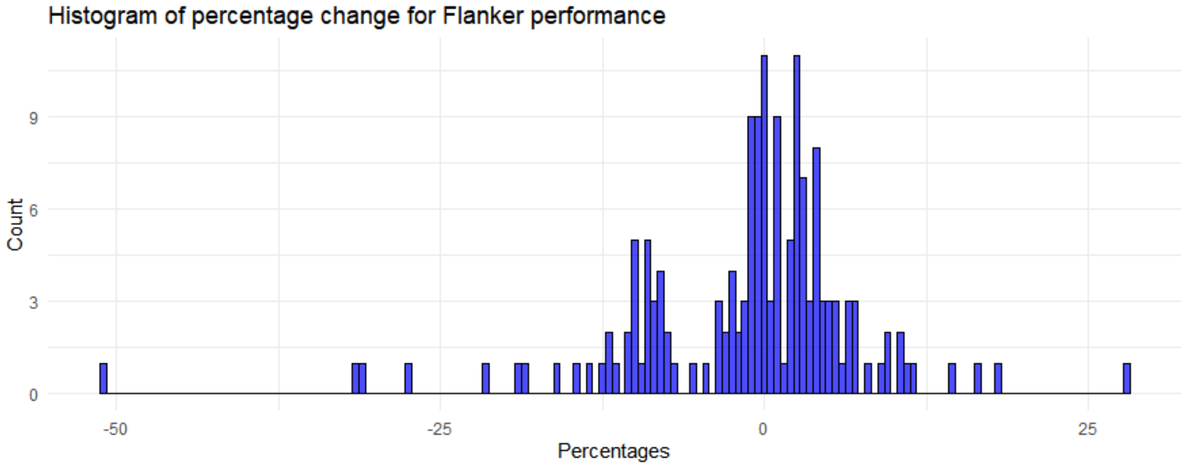


Figure 3.20: Percentage of changes over a year for Flanker game

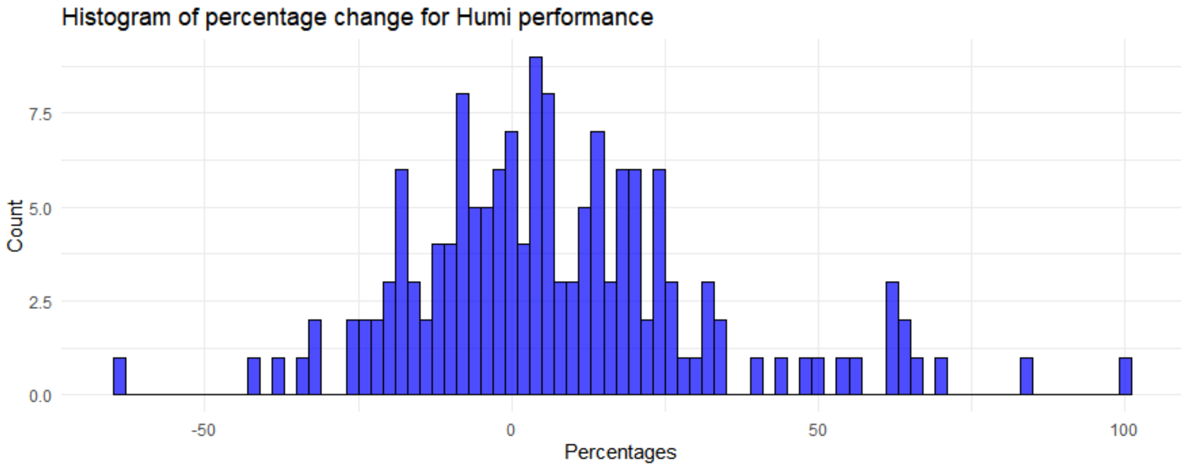


Figure 3.21: Percentage of changes over a year for Humi game

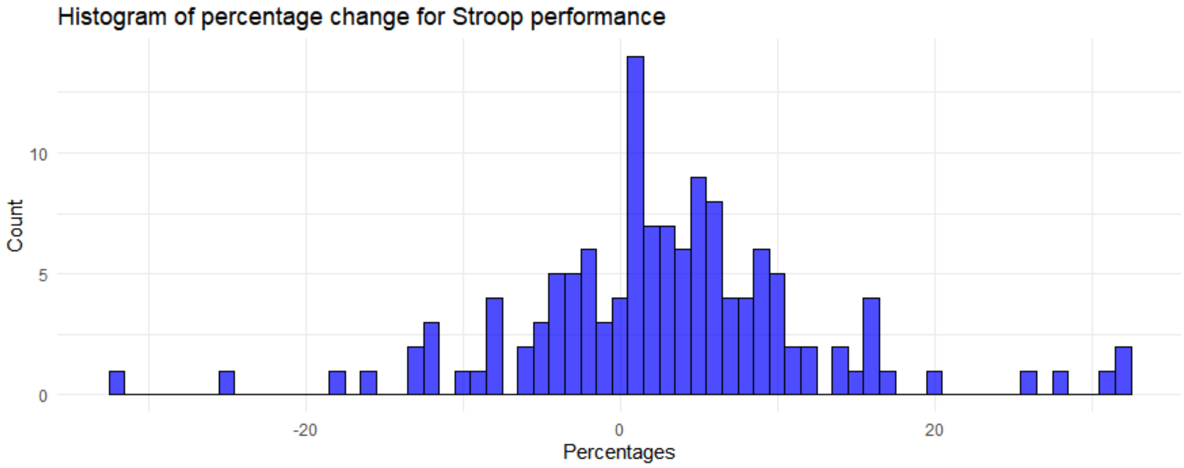


Figure 3.22: Percentage of changes over a year for Stroop game

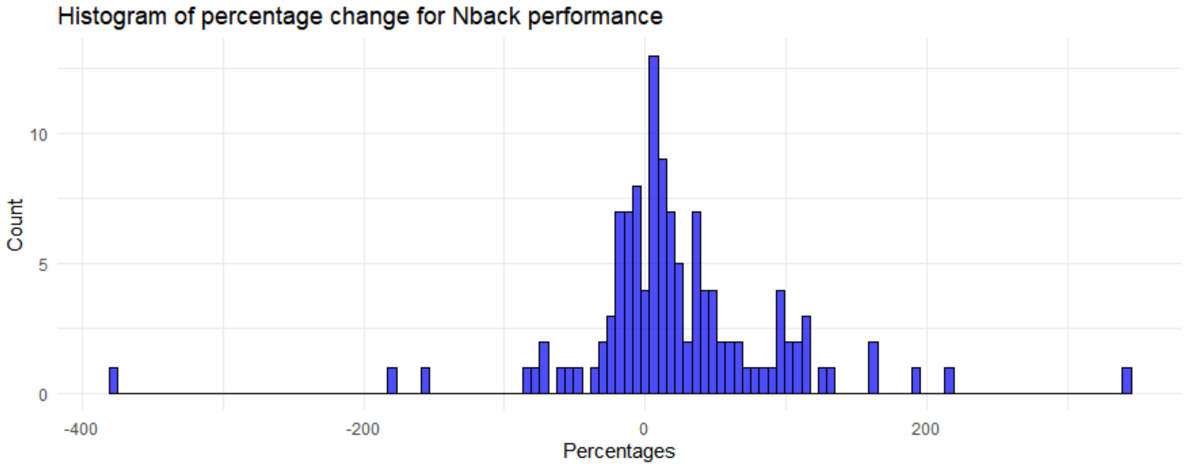


Figure 3.23: Percentage of changes over a year for Nback game

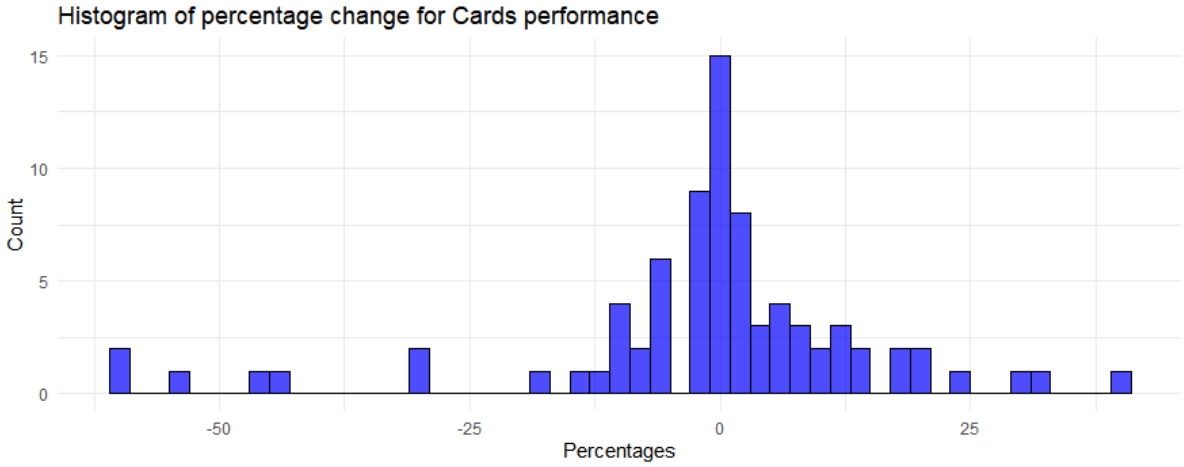


Figure 3.24: Percentage of changes over a year for Cards game

Games	n	Mean(SD)	Median	Min	Max
Flanker	157	-1.24(9.02)	0.14	-51.15	28.13
Humi	152	7.55(24.90)	4.76	-64.86	100.00
Stroop	131	2.92(9.66)	2.66	-31.55	32.13
Nback	121	22.11(74.95)	12.65	-379.28	345.00
Cards	79	-1.42(17.58)	0.00	-60.00	40.00

Table 3.3: Statistics for percentage of changes over time in games

Concluded Remarks

We conclude that brain volumes provide valuable insights into the progression of frontotemporal dementia, as they effectively distinguish patients with different CDR and NACC measurements. Given this, evaluating the relationship between brain volumes and cognitive game performance is a logical step. Our findings indicate a mild positive correlation between cognitive game performance and brain volumes, ranging from 0.3 to 0.6. Additionally, age is negatively correlated with both cognitive games and brain volumes, suggesting that cognitive games may indeed capture the decline in cognitive function over time. Another notable finding is the statistically significant difference across genders. Furthermore, cognitive games demonstrate the ability to differentiate between asymptomatic and healthy individuals based on CDR and NACC measurements, aligning with the results of previous studies.[2] Lastly, both brain volumes and game performance exhibit changes over time, with game performance showing greater variability. In the next section, we will apply Mixed-Effects Models and Structural Equation Modeling to test hypotheses based on these findings and further assess the relationship between cognitive games' performance and brain volumes.

Chapter 4

Statistical Analysis

4.1 Models

4.1.1 Multiple Linear Regression

The simplest model to build is a multiple linear regression model. By developing a separate model for each response variable, we can observe that the games are indeed related to the medical (laboratory) tests. The models take the following form:

$$Y = \beta_0 + \beta_1 * Age + \beta_2 * Time + \beta_3 * GamePerformance + \epsilon_0$$

However, fitting this model on longitudinal data, creates some crucial issues. The major problem we face is the assumption of independence in the residuals, which does not hold true in our dataset. By using a multiple regression model, we essentially pool all the data, which treats repeated measures for the same individual as though they were independent observations. This approach ignores the fact that measurements taken from the same participant over time are correlated. As a result, it's as if each time point for a subject is treated as a separate individual.

Applying this type of model to longitudinal data can lead to biased standard errors, rendering significance tests invalid. Additionally, there are unobserved, time-invariant variables that influence the outcomes of diagnostic tests beyond our predictors. These might include factors like a participant's IQ or other inherent characteristics that are not captured by our model. The pooled OLS model cannot account for the effects of these unobserved variables, leading to unexplained variability in our results.

4.1.2 Mixed-Effects Models

A solution to the problem of correlated residuals is the use of mixed-effects models. These models address this issue by incorporating random effects and, if necessary, specifying a correlation structure in the residuals. Random effects help explain the variability observed between individuals. In addition to fixed effects, which describe population-level trends similar to multiple regression, mixed-effects models introduce random effect parameters that account for individual differences. Why is this important? Why not rely solely on population averages? In many cases, factors influencing individuals may not be explicitly included in the model but still affect outcomes differently for each person. For instance, in clinical trials, patient-specific responses are often of interest because unmeasured individual characteristics can distort the estimated effects at the population level, potentially masking true relationships. Including random intercepts helps address this issue by capturing differences in baseline levels between subjects. Another key advantage of incorporating random intercepts is improved model fit. These models not only account for baseline differences but can also explain variability in slopes for a given variable. Additionally, they address correlated residuals, leading to more accurate standard errors and p-values. A random intercept model fits lines with different baselines but assumes constant slopes across participants. In contrast, adding random slopes makes the model more flexible, allowing individual slopes to vary. Once between-subject variability is accounted for, within-subject variability is captured through the residual variance.

Fitting the model in ALL-FTD dataset

We begin by fitting a random intercept mixed-effects model for each game performance variable, using left frontal lobe volume as the dependent variable. Additionally, we control for baseline age and semester, allowing us to gain insights into baseline differences among participants as well as the fixed effects of age, semester, and game performance. The model is estimated using maximum likelihood, assuming that missing data follow a Missing at Random (MAR) mechanism. Thus, the random intercept mixed model is structured as follows:

$$Y_{ij} = \beta_0 + \beta_1 * Age_j + \beta_2 * Semester_{ij} + \beta_3 * GamePerformance_{ij} + b_{0j} + \epsilon_{ij}$$

where:

- Y_{ij} denotes the brain region volume for subject j at time i,
- $GamePerformance_{ij}$ represents the mobile game score for subject j at time i,

- $Semester_{ij}$ represents the semester i for subject j for $i = 0, 1, 2, ...,$
- β_0, β_1, β_2, and β_3 are fixed effect coefficients (intercept, effect of baseline age, effect of time, and effect of game score, respectively),
- $b_{0j} \sim N(0, \sigma^2_{b_0})$ represents the random intercept for subject j,
- $\epsilon_{ij} \sim N(0, \sigma^2)$ denotes residual errors.

We can assess the significance of the random intercept by comparing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for models with and without the random intercept. Additionally, the intraclass correlation coefficient (ICC) provides insight into the proportion of total variance attributable to between-subject differences. If a substantial portion of the variance is due to the random intercept, we can infer its significance. After evaluating baseline between-subject variability, it is worth exploring whether there are differences in slopes across participants by fitting a random slope model. By introducing a random slope for cognitive game performance, we can assess whether the relationship between changes in game performance and changes in brain volume varies across individuals. If the random slope for game scores is statistically significant, this would indicate that the association between game performance and brain volume over time differs among subjects. The model capturing this variability is structured as follows:

$$Y_{ij} = \beta_0 + \beta_1 * Age_j + \beta_2 * Semester_{ij} + (\beta_3 + b_{1j}) * GamePerformance_{ij} + b_{0j} + \epsilon_{ij},$$

where b_{1j} corresponds to the random slope for the game performance

However, as we mentioned before the dataset contains 206 participants with only one measurement in brain volumes, 71 subjects with two time points and 2 with three. Therefore, it is not possible to fit random slope without having sufficient amount of at least 3 time points.

Linear Interpolation

It is essential to collect more longitudinal data to test the significance of the random slope assumption. However, since this is not feasible in the short term, we can make a stricter assumption about the timing of brain volume measurements. As previously mentioned, brain volumes are measured annually, while game performances are assessed every six months. To bridge this gap, we can impute the missing brain volume values using linear interpolation. This approach, while strong, may still provide some insight into the between-subject variability of slopes. That said, the results should be interpreted with caution, as linear interpolation is not always a valid assumption. To mitigate potential biases, we introduced random error into the imputed values, adding noise to address correlation bias. Some studies suggest that brain volume declines linearly over time, but this has not been universally proven. [6] Given the variability in dementia severity among participants, different stages of the disease may lead to distinct patterns of brain atrophy. Nonetheless, we proceeded with imputing missing values under the assumption of linear interpolation, carefully interpreting the results. With these imputed values, we were able to fit a mixed model with a random slope, but only for the Flanker game. It was not possible to estimate random slopes for the other games due to convergence issues stemming from missing data. While multiple imputation could be explored to address these missing values, layering additional assumptions onto an already strong assumption about brain volume trends would likely render the results even more unreliable. Thus, we focused on building a random slope model for Flanker to assess whether game performance can help explain differences in brain volume slopes between individuals.

Robust Linear Mixed-Effects Model

Another challenge in the analysis was the assumptions underlying mixed models, particularly the assumption of normality. Given that we are working with real-world data from mobile games and diverse participants, inconsistencies and outliers are inevitable. Real-life data rarely conform perfectly to statistical assumptions. To address this, we implemented a robust linear mixed-effects model, which provides more reliable and unbiased estimates by adjusting the influence of individual data points based on their residuals. This method assigns lower weights to observations that deviate significantly from the model's expectations. However, the implementation of robust mixed-effects modeling in R is currently limited to models with random intercepts, meaning we could not apply it to a mixed model with a random slope. Therefore, we present results for the robust linear mixed-effects model with a random intercept.

4.1.3 Bivariate Latent Change Score Model

The limitations discussed in the previous chapter regarding mixed-effects models are significant, preventing us from drawing meaningful conclusions about the impact of game performance on brain volumes over time. The inadequacy of data across time points and the strong assumption of linear interpolation caution us to interpret the results carefully. A model that relies solely on existing data may provide more reliable insights. The Bivariate Latent Change Score Model is one such approach that could yield better results. This model integrates a measurement model with a latent change score model, allowing us to examine the relationship between game performance and brain volumes over time. Specifically, we define a latent variable representing cognitive function—an abstract concept that cannot be directly measured. The cognitive games in our dataset attempt to assess this function, making them suitable indicators for our latent construct. By using observed game performances as indicators of cognitive function, we reduce the measurement error inherent in individual games. The measurement model validates whether these games accurately represent cognitive function, which is crucial given that frontotemporal dementia leads to brain atrophy, affecting problem-solving, attention, and decision-making abilities. If game performances can effectively measure cognitive function, they may provide valuable insights into the progression of the disease. The combination of the measurement model with the latent change score model allows us to test this hypothesis and assess its significance. In Figure 19, we present the Bivariate Latent Change Score Model. We selected three games—Flanker, Humi, and Stroop—as indicators of the cognitive function latent variable. These games are the most representative for this analysis. In contrast, N-back and Card Shuffle were excluded due to their lower correlation with brain volumes and high number of missing values, which could lead to biased and unreliable estimates. Including all available games would also be problematic, as an excessive number of indicators for a single latent variable can cause convergence issues and poor model fit. Additionally, we standardized brain volumes to mitigate potential convergence problems and ensure proper model estimation.

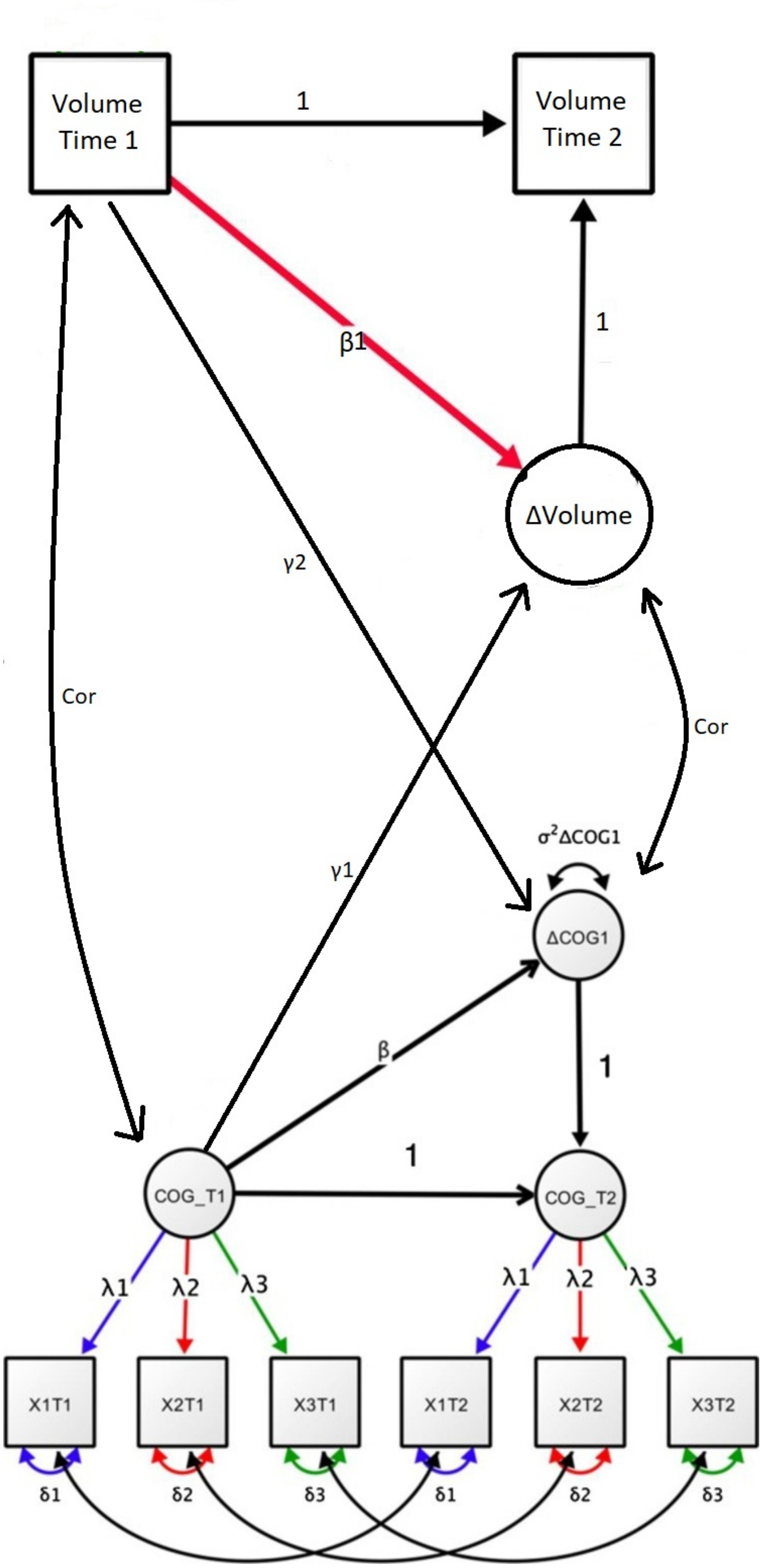


Figure 4.1: Latent Change score model for brain volume.

4.2 Results

4.2.1 Mixed model with Random Intercept

The mixed model with a random intercept indicates a statistically significant variance between subjects in baseline brain volumes ($ICC = 0.97$). However, since the dataset includes at most two time points per subject, it is reasonable for this coefficient to be high. With only two time points, there is limited within-subject variability, which results in a lower proportion of variance being captured by residuals. Despite this, incorporating random effects remains beneficial, as each individual has unique characteristics. If we aim to explore the factors contributing to differences in brain volumes in future research, including random effects is essential. Additionally, the assumption of independent residuals is violated, which is accounted for by incorporating random effects. Furthermore, model selection criteria such as AIC and BIC can be used to compare models with and without a random intercept. In Table 8, the AIC and BIC values for the model using the Flanker game as an independent variable favor the inclusion of a random intercept. Similar results are observed for models incorporating other cognitive games.

Flanker Model	Random Intercept	Without Random Intercept
AIC	-3222	-3145
BIC	-3199	-3125

Table 4.1: Model comparison with and without random intercept

After testing about the significance of random intercept, we can fit the model and observe the results. In Table 9, the coefficients of fixed and random effects are presented. Below, we demonstrate the interpretation of the coefficients for the Flanker Model. The same interpretation applies for the other models too.

Intercept: For an 18-year-old individual with an average Flanker game score, the left frontal lobe volume constitutes on average 3.57% of the total intracranial volume during the baseline semester.

Baseline Age: Controlling for other variables, a year older person has on average decrease of 0.02% in the relative size of the brain region (brain region volume normalized by intracranial brain volume).

Semester: Controlling for other variables, after 6 months, a person has on average a decrease of -0.016% in the relative size of the brain region.

Flanker Game: Controlling for other variables, a unit increased in the score of the game, we have on average a 0.009% increase in the relative size of the brain region.

Model	Intercept	Baseline Age	Semester	Game	Random Intercept Standard Deviation	Residual Standard Deviation
Flanker	3.57%	-0.02%	-0.016%	0.009%	0.278%	0.043%
Humi	3.56%	-0.02%	-0.016%	0.023%	0.279%	0.045%
Cards	3.56%	-0.02%	-0.016%	0.005%	0.276%	0.028%
Nback	3.56%	-0.018%	-0.015%	0.002%	0.263%	0.042%
Stroop	3.59%	-0.015%	-0.016%	-0.018%	0.260%	0.039%

Table 4.2: Fixed and Random coefficients of models

Figures 19, 20, and 21 illustrate the confidence intervals for the fixed effects coefficients. The baseline Age and Semester are statistically different from zero for all the models. For the games' performances only the Cards Shuffle is statistically different from zero.

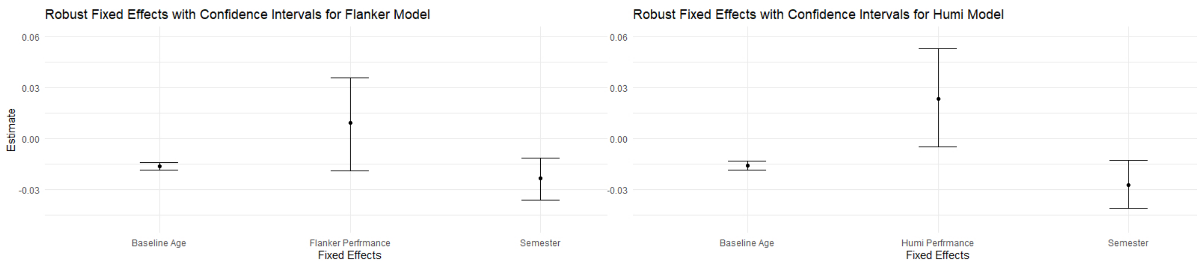


Figure 4.2: Confidence intervals for fixed effects in Flanker and Humi models

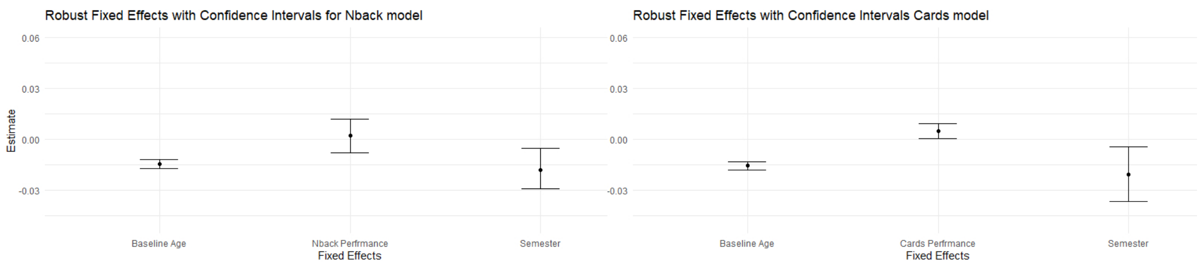


Figure 4.3: Confidence intervals for fixed effects in Nback and Cards models

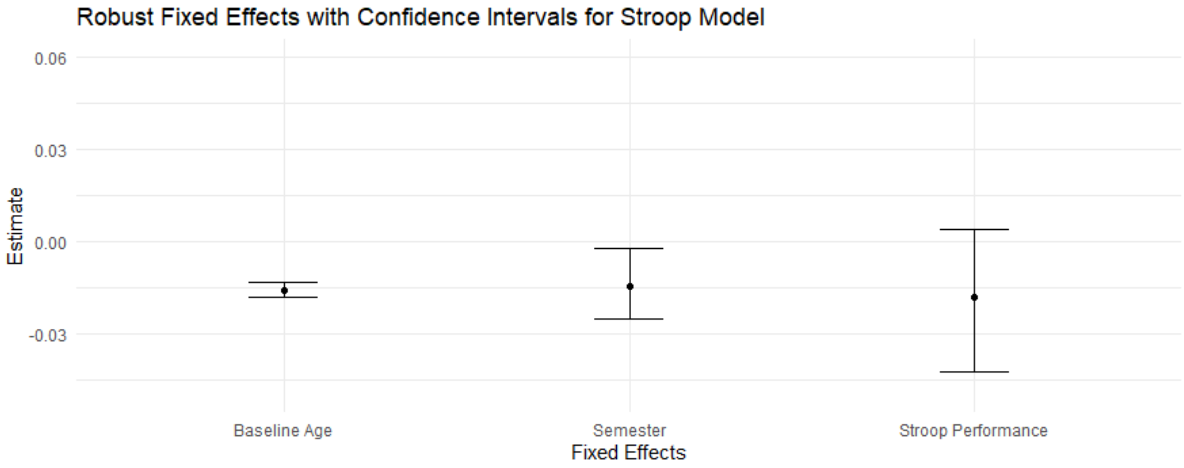


Figure 4.4: Confidence intervals for fixed effects in Stroop model

4.2.2 Mixed-Effects Model with Random Slope

We will examine now whether there is variation in the slopes of subjects for the Flanker game. As we mentioned before, the dataset contains at most two time points for each subject and we linearly interpolated these two time points to obtain three time points and have the opportunity to fit a random slope. Moreover, the results should be interpreted with cautious because of the strong assumption of linear interpolation. The standard deviation of the random slope is approximately zero, indicating that the slopes of subjects do not vary significantly. In addition, AIC and BIC favor the mixed model with random intercept only(Table 10).

Flanker Model	Random Slope	Without Random Slope
AIC	-3877	-3881
BIC	-3845	-3857

Table 4.3: Model comparison with and without random slope

4.2.3 Bivariate Latent Change Score Model

Bivariate Latent Change Score model gives us the opportunity to directly measure some interesting results between games' performance and brain volume, which were not eligible in mixed-effects models. In the previous section, we have constructed the ideal model for the analysis. After that, the initial task is to evaluate the model's fit. Table 11 indicates the fit indices of the model, which meet the recommended thresholds.

Fit Index	Value
CFI	0.981
TLI	0.977
RMSEA	0.051
SRMR	0.040

Table 4.4: Fit Indices of the model

We conclude that the fit indices support the model, allowing us to proceed with inference. The cognitive construct(COG_1) and the observed volume brain($Volume_1$) aat time 1 have positive relationships with brain volume changes($\gamma_1 = 0.095$ and $\beta_1 = 0.085$ respectively, $pvalue < .05$). This suggests that individuals with better game performance at baseline tend to experience smaller negative changes in brain volume over time, which aligns with expectations. Healthy individuals generally perform better in cognitive games, indicating stronger cognitive function, and thus exhibit smaller changes in brain volume over time. Conversely, individuals with dementia are expected to perform worse in the games and experience greater cognitive decline, leading to larger brain volume changes over time. Additionally, individuals with larger brain volumes at baseline tend to show less change in brain volume over time, which is also supported by the model's findings.

The relationship between $Volume_1$ and changes in cognitive function (ΔCOG) is positive ($\gamma_2 = 0.093$, $pvalue > .05$). meaning that individuals with higher brain volume at baseline tend to show greater changes in game performance over time. This could be attributed to the dataset including both healthy and symptomatic individuals. Healthy individuals may improve their performance over time, while symptomatic individuals may decline. When these trends are aggregated, the net effect of brain volume appears positive. However, the model does not reject the hypothesis that this effect is different from zero. Similarly, the association between Cog_1 and ΔCOG is not statistically significant.

The correlation between $Volume_1$ and COG_1 is positive ($Cor = 0.624$, $pvalue < .05$), indicating that individuals with higher brain volumes tend to perform better in cognitive games and exhibit stronger cognitive function. However, we cannot reject the hypothesis that the correlation between changes in the latent constructs (ΔCOG and $\Delta Volume$) is zero ($Cor = -0.110$, $pvalue > .05$)

Furthermore, the variances of the changes in the latent constructs(ΔCOG and $\Delta Volume$)

are statistically different from zero ($\sigma^2 \Delta COG = 0.066$, $\sigma^2 \Delta Volume = 0.065$ with $pvalue < .05$). This indicates that there are significant individual differences in the changes in both game performance and brain volume over time.

Measurement Invariance

As mentioned earlier, there is a bias toward males in the number of symptomatic participants, with more symptomatic individuals in the male group than in the female group. Including this categorical variable in the model may introduce bias in the coefficients, leading to unreliable estimates. To test whether the latent constructs measure the same concept without bias in the structural equation model, we impose constraints on certain parameters for both males and females and assess the model fit. To determine whether the construct is measured equivalently across genders, we must establish scalar invariance. This requires constraining both the factor loadings and the intercepts of the indicators (cognitive games) to be the same across genders. We then compare this model to the unconstrained metric invariance model. If we fail to reject the difference between the metric invariance model and the configural invariance model (which allows factor loadings and intercepts to vary freely), we can assume that measurement properties are partially consistent across genders. In Table 12, we observe that we cannot reject the hypothesis that configural invariance and metric invariance models are different. However, we do reject the hypothesis that metric and scalar invariance models are equivalent. This means we cannot assume that the latent construct is measured consistently across genders.

Invariance	Df	AIC	BIC	Chi-Squared	P-Value
Configural	42	3577	3744	87.71	
Metric	44	3574	3734	89.14	0.64
Scalar	47	3607	3756	128.01	<0.01

Table 4.5: Model Comparison for measurement invariance in gender

Summary

In this chapter, we performed Mixed-Effects Models and the Bivariate Latent Change Score Model. Overall, we found no evidence that changes in game performance are related to changes in brain volumes. However, the Mixed-Effects Models revealed that the random intercept is statistically significant, indicating variability in baseline measures between subjects. Additionally, we observed a statistically significant decline in brain volumes with age. After controlling for age, we could not reject the hypothesis that cognitive game performance has an effect different from zero, except in the case of the Cards Shuffle game. We attempted to fit a random slope after imputing missing six-month brain volume values, but the random slope was not statistically significant. Therefore, we cannot conclude that there is variability in brain volume trajectories based on Flanker game performance over time. However, these results may be unreliable due to the small sample size and the strong assumption of linear interpolation.

On the other hand, the Bivariate Latent Change Score Model provided valuable insights. Overall, there are inter-individual differences in brain volumes and cognitive game performance over time. Additionally, we found a strong positive correlation between the baseline latent construct of cognitive games and baseline brain volume, suggesting that, on average, higher game scores correspond to larger brain volumes. Lastly, the model suggests that cognitive game scores and brain volumes may offer meaningful insights into brain volume trajectories.

Chapter 5

Conclusion and Discussion

It would be beneficial to assess cognitive function and its decline in patients with Frontotemporal dementia, a disease that accelerates brain volume atrophy and negatively impacts patients' lives. MRI scans have traditionally been the primary objective method to track this decline, but they are time-consuming and costly, making it difficult to obtain a sufficient amount of data. This limitation poses challenges for clinical trials and patient monitoring. In contrast, mobile cognitive games, as presented in this study, offer a practical alternative by providing valuable information about the stage of dementia in patients. These games are easy to administer, convenient for patients, and simple to implement. In this study, we introduced a mixed-effects model and a bivariate latent change score model to examine the relationship between brain volumes and mobile game performance. The primary objective was to evaluate these games as potential biomarkers or as complementary tools alongside more established biomarkers. Our findings suggest a positive association between game performance and brain volumes in both models. Moreover, the bivariate latent change score model provided additional insights into the association between changes in game performance and changes in brain volume over time, an aspect that was difficult to capture using the mixed-effects model due to the limited amount of available data. Specifically, only two time points were available for brain volume measurements, restricting our ability to fully utilize mixed models and structural equation modeling, particularly in incorporating random slopes or drawing reliable conclusions about long-term trajectories.[17] Previous research has suggested that if time points are sufficiently spaced apart, even two measurements can yield more reliable results.[4] Nonetheless, while the bivariate latent change score model identified individual differences in both game performance and brain volume changes, these changes were not directly associated with each other, highlighting the need for further investigation.

There were several limitations in this study that future research should address. The collection of additional data would allow for the development of more refined measurement models and a more comprehensive representation of cognitive function derived from mobile games. Incorporating additional games could enhance the ability to construct complex latent variables that better capture the intricacies of brain function and provide a more accurate assessment of cognitive abilities. Instead of focusing solely on cognitive function as a latent construct, the availability of a broader range of serious games and their associated data would enable the creation of additional latent variables, such as memory, which is also affected by frontotemporal dementia. Furthermore, developing a latent variable for brain volumes that accounts for measurement error from MRI scans would improve the reliability of the results. If both cognitive function and memory were represented as latent constructs derived from serious games and brain volumes were adjusted for measurement error, the association between serious games and brain volumes could be assessed more accurately. Beyond data collection, obtaining repeated measurements over time would significantly enhance the construction of mixed-effects models, allowing for a more detailed examination of longitudinal relationships and the incorporation of random slopes. With more time points, it would also be possible to develop a bivariate dual latent change score model, introducing an additional latent construct that captures overall trends and facilitates comparisons of latent variables across different periods. This would provide a more comprehensive understanding of the relationship between game performance and brain volumes over time. In addition, an alternative approach could involve Bayesian modeling, incorporating uninformative or, if possible, informative priors. Bayesian methods in mixed models and structural equation modeling offer advantages in addressing missing data issues, even in cases where data are missing not at random. They are also effective in small sample situations and provide posterior distributions for each parameter, making them particularly useful when dealing with identification or convergence problems or difficulties in fitting random slopes.[14] Given these advantages, Bayesian modeling could be a promising direction for future research in this field.

Bibliography

[1] A., B. [2024], 'Digital biomarkers in parkinson's disease.', *Adv Clin Chem.* .

[2] Adam M. Staffaroni, P. et al. [2024], 'Reliability and validity of smartphone cognitive testing for frontotemporal lobar degeneration'.

[3] Allison, P. D. [2002], *Missing Data*, 1st edn, SAGE Publications, Inc.

[4] Andreas M. Brandmaier, Ulman Lindenberger, E. M. M. [2024], 'Optimal two-time point longitudinal models for estimating individual-level change: Asymptotic insights and practical implications'.

[5] Bang, J., S. S. . M. B. L. [2015], 'Frontotemporal dementia'.

[6] C. E. Coffey, M. et al. [1992], 'Quantitative cerebral anatomy of the aging human brain'.

[7] Convery, R. et al. [2024], 'Concurrent validity, test-retest reliability, and normative properties of the ignite app: a cognitive assessment for frontotemporal dementia'.

[8] Coughlan, G. et al. [2019], 'Toward personalized cognitive diagnostics of at-genetic-risk alzheimer's disease'.

[9] Daire Hooper, J. C. and Mullen, M. R. [2007], 'Structural equation modelling: Guidelines for determining model fit', *The Electronic Journal of Business Research Methods* .

[10] Davidian, M. [2005], *Applied Longitudinal Data Analysis*, 1st edn, North Carolina State University.

[11] Greaves, C. V., . R. J. D. [2019], 'An update on genetic frontotemporal dementia'.

[12] K, G. et al. [2021], 'Dissecting digital card games to yield digital biomarkers for the assessment of mild cognitive impairment: Methodological approach and exploratory study', *JMIR Serious Games* .

[13] Kievit, R. A. et al. [2017], 'Developmental cognitive neuroscience using latent change score models: A tutorial and applications', *https://pmc.ncbi.nlm.nih.gov/articles/PMC6614039/* .

[14] Ouyang, J. et al. [2025], 'Bivariate latent change score models in small sample contexts: Comparison of maximum likelihood and bayesian estimation'.

[15] Putnick DL, B. M. [2016], 'Measurement invariance conventions and reporting: The state of the art and future directions for psychological research', *Dev Rev.* .

[16] Roback, P. and Legler, J. [2021], *Beyond Multiple Linear Regression*, 1st edn, Chapman Hall/CRC.

[17] Sam Parsons, E. M. M. [2024], 'Limitations of two time point data for understanding individual differences in longitudinal modeling — what can difference reveal about change?'.

[18] Schober P, V. T. [2018], 'Repeated measures designs and analysis of longitudinal data: If at first you do not succeed-try, try again.', *Anesth Analg.* .

[19] Sturgis, P. P. [2016], 'Confirmatory factor analysis'.

[20] Taylor, J. C. et al. [2023], 'Feasibility and acceptability of remote smartphone cognitive testing in frontotemporal dementia research'.

[21] van Buuren, S. [2018], *Flexible Imputation of Missing Dara*, 2nd edn, Chapman Hall/CRC.

[22] Wang, C. S.-M. et al. [2022], 'Using self-administered game-based cognitive assessment to screen for degenerative dementia: A pilot study'.

[23] Werner, P. C. and Schermelleh-Engel, P. D. K. [2009], 'Structural equation modeling: Advantages, challenges, and problems'.