

Assessing the Effect of tACS with Cognitive Exercise in Dementia Population: A Statistical Perspective on Crossover Trial Analysis

by

Elham Mahdavi Asiabar

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of
The University of Manitoba
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Biomedical Engineering Program

University of Manitoba

Winnipeg, Manitoba

Copyright © 2026 by Elham Mahdavi Asiabar

Abstract

Alzheimer’s disease (AD) has limited treatment options, motivating growing interest in non-invasive approaches such as gamma-frequency transcranial alternating current stimulation (tACS) combined with cognitive exercises (CE). This thesis investigates the cognitive effects of tACS+CE using a randomized, double-blind 2×2 crossover design in individuals with dementia. Participants completed real and sham stimulation across two four-week treatment periods separated by a 2–5-month washout, with baseline and post-treatment ADAS-Cog assessments collected in each period. The crossover design allows each participant to serve as their own control while enabling evaluation of treatment, period, sequence, and carryover effects.

Treatment effects were examined using classical frequentist, bootstrap, and Bayesian methods. Classical analyses, including paired t -tests, crossover ANOVA, and ANCOVA with baseline adjustment, showed no statistically significant difference between real and sham tACS, with baseline cognitive status emerging as the strongest predictor of post-treatment performance. Bootstrap analyses confirmed these findings, with resampled confidence intervals consistently spanning zero and null-based bootstrap tests indicating no evidence of a treatment effect. In contrast, the Bayesian analysis revealed moderate evidence in favor of a beneficial treatment effect when results were averaged across competing models, yielding an unconditional posterior probability $\mathbb{P}(\tau > 0 \mid Y) = 0.6117$, indicating that, given the observed data, there is approximately a 61% probability that the active intervention improves cognitive performance relative to sham treatment. Collectively, these results indicate that while classical and bootstrap methods provide no evidence of short-term cognitive benefit in this small-sample crossover study, the Bayesian framework highlights sensitivity to modeling assumptions and offers a complementary, probabilistic perspective on treatment efficacy. This thesis underscores the importance of combining multiple inferential approaches when evaluating non-pharmacological interventions in dementia research.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Mohammad Jafari Jozani, for his continuous guidance, patience, and invaluable support throughout my Master's studies. His expertise in statistical methodology and his thoughtful feedback were instrumental in shaping this thesis. I am particularly grateful for his patience in explaining complex statistical concepts and for helping me develop a more systematic and rigorous approach to statistical analysis. His emphasis on mathematical clarity and methodological precision has had a lasting impact on my academic development. Beyond academic supervision, he broadened my perspective and influenced the way I understand statistics and scientific inquiry.

I would also like to extend my sincere appreciation to my co-supervisor, Dr. Zahra Moussavi, for giving me the opportunity to join her research group and for her support throughout this work. Her scientific insight and interdisciplinary perspective greatly enriched this research. Through her mentorship, I developed a deeper appreciation for research that bridges engineering innovation with meaningful clinical needs. Her strong commitment to patient-centered research and dedication to improving cognitive health have been truly inspiring. As a biomedical engineer, I deeply admire her leadership and the example she sets as a researcher and mentor.

My sincere thanks go to my committee members, Dr. Ke Peng and Dr. Sumeet Kalia, for their time, careful review, and constructive feedback. Their valuable suggestions significantly enhanced the quality and clarity of this thesis.

I would also like to acknowledge the University of Manitoba for providing an excellent academic environment and the resources necessary to complete this research.

Last but not least, I am deeply grateful to my family and friends, whose constant love and encouragement have been the foundation of my strength throughout this journey.

Pursuing my studies away from my home country has been both challenging and transformative, and their trust in me has carried me through moments of uncertainty. No matter the distance, their support has remained a steady source of motivation and reassurance.

Dedication

To my dear uncle, Fardin,

who believed in me long before I believed in myself.

Your encouragement, love, and unwavering support have guided me since childhood.

Though you are no longer here, your presence lives in my heart,

and this achievement is, in many ways, yours.

Contents

Abstract	ii
Acknowledgments	iii
List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Cross-over Design Study	1
1.2 Alzheimer’s Disease	3
1.3 Existing Treatments for AD	4
1.3.1 Non-Invasive Brain Stimulation (NIBS) Techniques	5
1.3.2 Transcranial Alternating Current Stimulation (tACS)	6
1.3.3 Cognitive Exercises (CE)	8
1.3.4 Single and multi-Sessions Gamma-tACS Studies	9
1.4 TACS Crossover Trial Study Design and Participants	10

1.5	Research Objectives	17
1.6	Thesis Organization	18
2	Frequentist analyses of 2×2 crossover studies	19
2.1	Introduction	20
2.2	Data Preparation and Descriptive Statistics	21
2.2.1	Dataset Description	22
2.2.2	Computation of Within-Period Change Scores	22
2.2.3	Treatment and Period Labels	23
2.2.4	Descriptive Statistics	23
2.2.5	Initial Observations from the Data	26
2.3	Assumption Checking	27
2.3.1	Assessment of Normality	27
2.3.2	Homogeneity of Variances	29
2.3.3	Baseline Equality Between Sequences	30
2.4	Testing for Design-Specific Effects	31
2.4.1	Sequence Effect	31
2.4.2	Period Effect	32
2.4.3	Carryover Effect	33
2.5	Treatment Effect Analysis	33
2.5.1	Paired t -test for Within-Subject Treatment Effect	34
2.5.2	Treatment Comparison with Two-Sample t -test	35

2.5.3	Repeated Measures ANOVA	36
2.5.4	ANCOVA Adjusting for Baseline Differences	38
2.6	Conclusion	40
3	Bootstrap analyses of 2×2 crossover studies	41
3.1	Introduction	41
3.2	A high level introduction to bootstrap	43
3.2.1	Bootstrap as an approximation to the sampling distribution	44
3.2.2	Bootstrap hypothesis testing	45
3.2.3	Variants of the bootstrap and relevance for this study	47
3.3	Bootstrap Treatment Comparison	48
3.3.1	Nonparametric subject-level bootstrap	49
3.3.2	Bootstrap standard error, confidence interval, and test	49
3.3.3	Bootstrap Crossover ANOVA	51
3.3.4	Cluster (subject-level) bootstrap scheme	51
3.3.5	Bootstrap inference for treatment and period effects	52
3.3.6	Bootstrap ANCOVA with Baseline Adjustment	53
3.3.7	Nonparametric subject-level bootstrap	54
3.3.8	Residual cluster bootstrap under the null	55
3.4	Conclusion	57
4	Bayesian analyses of 2×2 crossover studies	58

4.1	Introduction	58
4.2	Bayesian Framework for Crossover Study	61
4.2.1	Models in the two-period crossover with baselines	61
4.2.2	Multivariate Normal Framework for Longitudinal Data	63
4.2.3	Prior distribution	65
4.2.4	Likelihood function	66
4.2.5	Posterior Distribution	68
4.2.6	Bayes Factors	82
4.3	Conclusion	96
5	Discussion and Conclusion	97
5.1	Discussion	98
5.2	Conclusions	100
5.3	Limitations	102
5.4	Future Research Directions	103
5.5	Final Remarks	104

List of Tables

1.1	Descriptive statistics summarizing the demographic characteristics of study participants stratified by group.	12
2.1	Descriptive statistics for ADAS–Cog scores by sequence	23
4.1	Cell mean structures under various crossover models.	62
4.2	Conditional posterior distributions in model $M_i : N(\alpha, \beta m \sigma^2)$	77
4.3	Summary statistics for ADAS-Cog data	81
4.4	Posterior Probability of Positive Treatment Effect Under Each Model	81
4.5	Interpretation of Bayes factors in support of M_0 (i.e., BF_{01})	83
4.6	Bayes factors comparing models	95
4.7	Posterior model probabilities for ADAS-Cog data	95

List of Figures

1.1	The structure of a 2×2 or AB/BA crossover design with washout period. . .	2
1.2	Illustration of the crossover study framework used in this thesis	11
1.3	Soterix Medical 1×1 transcranial alternating current stimulation (tACS) device and EASYstrap head strap used for electrode placement. (Source: Soterix Medical Inc., reproduced for educational purposes; permission not required.)	15
2.1	Boxplots of ADAS–Cog Scores and Within-Period Change Scores by Sequence AB	24
2.2	Boxplots of ADAS–Cog Scores and Within-Period Change Scores by Sequence BA	24
2.3	Histograms of ADAS–Cog change scores across sequences (AB and BA). . .	25
2.4	Q–Q plots of baseline ADAS–Cog scores for Sequence AB (left) and Sequence BA (right).	29
4.1	Posterior Distribution of Treatment Effect (τ) under four models, each imposing different constraints on the carryover (θ) and interaction (λ) effects .	80

Chapter 1

Introduction

Summary. In this chapter, we study crossover designs as a suitable framework for Alzheimer’s disease (AD) research. We first provide a brief survey of existing treatments and non-pharmacological options for AD, including non-invasive brain stimulation, transcranial alternating current stimulation (tACS), and cognitive exercises (CE). We then describe a randomized, double-blind 2×2 crossover study of gamma-frequency tACS combined with CE versus sham stimulation, comprising two four-week phases separated by a 2–5-month washout period, with pre- and post-treatment assessments using the Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog). We provide descriptive statistics and outline the research objectives. The chapter concludes by describing the thesis structure and the statistical methods that will be used to study treatment, period, and carryover effects with baseline adjustment in our crossover study.

1.1 Cross-over Design Study

Crossover designs provide a rigorous and efficient framework for evaluating therapeutic interventions, particularly in chronic and non-curable conditions such as asthma, hypertension, epilepsy, gastrointestinal reflux, arthritis, and migraine [1]. Their defining advantage is that

each participant receives multiple treatments over time and therefore serves as their own control. This within-subject structure substantially reduces between-individual variability, increases statistical power, and allows for smaller sample sizes compared to parallel-group designs [2]. In a standard crossover trial, N subjects are assigned to receive t treatments over p periods, with assignments organized into $m \leq N$ unique sequences s_i , each of length p .

The simplest and most widely adopted structure is the AB/BA or 2×2 crossover design (Figure 1.1), where participants are randomized to receive treatment A followed by treatment B or the reverse. Because outcomes are repeatedly measured within each participant, this design efficiently controls for confounding variables and yields highly precise estimates of treatment effects. Crossover studies are most appropriate when the underlying condition is stable over time and when treatment effects are transient and reversible. They are not suitable for curative therapies or conditions expected to change rapidly, where initial treatment may permanently alter the subsequent response trajectory. [3].

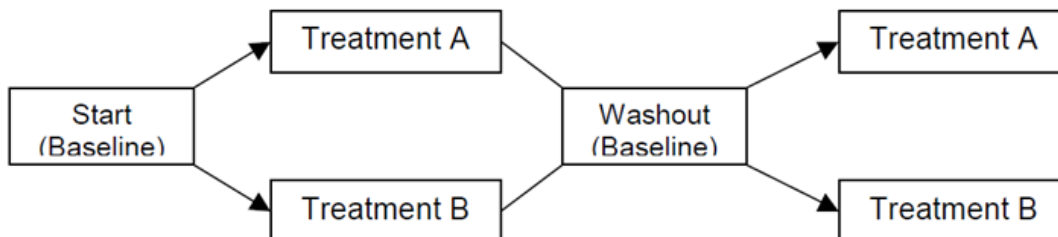


Figure 1.1: The structure of a 2×2 or AB/BA crossover design with washout period.

A central methodological challenge in crossover designs is the possibility of carryover effects, where the impact of a treatment persists into the next period and contaminates the response to the subsequent intervention [4]. To mitigate this, washout periods are typically inserted between treatment phases to allow participants to return to baseline clinical status. Even with washout, treatment-by-period interactions may still arise and can bias treatment comparisons. When strong carryover is suspected, analyses are sometimes restricted to first-period data, although this sacrifices the efficiency advantage of the crossover structure.

The literature on crossover methodology has expanded substantially, yet many published studies remain limited by design flaws, inappropriate modeling, or incomplete consideration of underlying assumptions [5]. A classical statistical model for the 2×2 design showed that it is impossible to estimate treatment, period, and carryover effects simultaneously and that valid within-patient treatment comparisons require ignoring carryover [6]. A commonly recommended approach is the two-stage testing procedure, where researchers first test for evidence of carryover and then combine both periods for treatment comparison only if no carryover is detected. However, later evaluations showed that this method can inflate Type I error rates and introduce bias, especially in small samples or when its underlying assumptions are not fully met [7].

In response, considerable attention has been devoted to exploring alternative crossover designs and analytical methods that mitigate these limitations. Several statistical investigations have evaluated design modifications intended to reduce susceptibility to carryover effects, such as extended washout periods, baseline adjustments, and sequences beyond two periods [8]. Nonetheless, many proposed solutions rely on the assumption that carryover lasts only one period and is unaffected by treatments administered later, an assumption increasingly regarded as overly simplistic, especially for interventions capable of inducing long-lasting physiological or behavioral changes [9]. When carefully designed and appropriately analyzed, however, crossover trials remain a powerful tool for assessing treatment efficacy in stable chronic conditions.

1.2 Alzheimer’s Disease

Due to the increase in life expectancy, it is projected that by 2050, approximately two billion individuals, almost 20% of the global population, will be aged 60 years or older [10]. This demographic shift is expected to lead to a rise in age-related chronic conditions and diseases.

One significant concern is cognitive decline, which can severely impact health and contribute to higher healthcare costs. Although mild cognitive decline may be a normal part of aging, more severe impairments in thinking and daily functioning can be early indicators of major neurodegenerative conditions such as dementia [11].

Dementia is a neurological disorder marked by the decline of multiple cognitive functions, including memory, communication, reasoning, and attention. Several types of dementia exist, including Alzheimer’s disease (AD), frontotemporal dementia (FTD), Lewy body dementia (LBD), vascular dementia (VaD), and mixed dementia (MxD). Among these, AD is the most prevalent, accounting for 60–80% of dementia cases globally [10].

Alzheimer’s disease is a progressive neurodegenerative disorder that gradually affects brain function, leading to a steady decline in memory and cognitive abilities. As the disease progresses, individuals lose the ability to perform daily tasks independently [12]. The impact of AD goes beyond normal aging, resulting in serious challenges for both patients and their caregivers, such as reduced quality of life and emotional burden. The symptoms of AD stem from neurodegeneration, the progressive loss of neurons in the brain [13]. Over time, this leads to a reduction in brain volume. As neuronal death becomes widespread, individuals experience worsening memory loss, and increasing difficulties with movement, communication, and cognitive functions [14].

1.3 Existing Treatments for AD

A persistent challenge in AD is the limited effectiveness of current medications in slowing or halting disease progression [15]. The structural and functional damage caused by AD is irreversible [16], and although some pharmacological treatments are available, they primarily serve to alleviate symptoms rather than halt or reverse the progression of the disease [17, 18]. Clinical trials for AD treatments have shown limited success overall, with only about 20%

of patients showing significant improvement and around 30% experiencing a temporary stabilization of symptoms [19, 20]. Current pharmacological treatments, such as cholinesterase inhibitors and N-methyl-D-aspartic acid (NMDA) receptor antagonists, provide only modest symptomatic relief and have limited long-term efficacy. Despite significant research efforts and therapeutic developments, there remains a critical need for innovative strategies to better understand and address the cognitive decline associated with AD. Due to the limited effectiveness of current drug treatments for AD, researchers have increasingly turned their attention to non-pharmacological approaches. These include lifestyle modifications, physical exercise, cognitive training, and non-invasive brain stimulation (NIBS) therapies as potential strategies to help manage cognitive decline in individuals with AD [21, 22, 23, 24, 25, 26]. Over the past several decades, extensive research has focused on investigating alterations in brain oscillations associated with age-related cognitive decline and Alzheimer’s disease [27].

1.3.1 Non-Invasive Brain Stimulation (NIBS) Techniques

In recent years, NIBS has attracted growing interest for its potential benefits in treating neurological disorders like dementia. NIBS involves several techniques, including transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), and tACS [22, 23, 24, 25, 26, 28]. These techniques work by either generating a magnetic field to induce electrical currents in the brain’s neocortex (TMS) or by delivering low-intensity electrical currents through the scalp (tDCS and tACS) that reach the neocortical areas. The choice of a specific brain stimulation method for studying cognitive enhancement in individuals with dementia depends on several factors, such as safety, effectiveness, and the goals of the research. In this thesis, transcranial alternating current stimulation (tACS) combined with cognitive exercises was selected as the primary intervention approach.

1.3.2 Transcranial Alternating Current Stimulation (tACS)

tACS is a non-invasive brain stimulation technique that applies low-amplitude alternating electrical currents at specific frequencies to modulate brain oscillations and influence cognitive functions [29, 30].

Transcranial alternating current stimulation (tACS) represents a valuable tool for investigating the relationship between frequency-specific brain oscillations and cognitive functions, as it delivers a sinusoidal current that oscillates at a defined frequency. By directly interacting with and entraining ongoing neuronal activity within targeted cortical networks, tACS can modulate task-relevant oscillatory dynamics. This frequency-specific synchronization allows researchers to infer causal links between neural rhythms and cognitive processes. Moreover, by selectively enhancing or altering oscillatory activity in specific frequency bands, tACS holds promise as a method for modulating cognitive performance through targeted influence on large-scale neural networks [31].

Unlike tDCS, which alters neuronal excitability through shifts in membrane potential (depolarization or hyperpolarization), tACS functions by synchronizing neuronal firing with an externally applied frequency. In tACS, one electrode serves as the anode and the other as the cathode during one half of the oscillatory cycle, with the current amplitude following the shape of a half sine wave. This polarity then reverses during the second half of the cycle. While both tDCS and tACS are being investigated as potential treatments for cognitive decline in dementia, emerging evidence indicates that tACS may offer greater efficacy compared to tDCS [32]. tACS administered at a frequency of 40 Hz has been reported to improve cognitive performance, likely through the enhancement of beta-band activity. This effect was found to be more pronounced than that observed with either sham stimulation or tDCS.

Depending on the study objectives, tACS can be administered to different scalp regions, such as the primary motor cortex, premotor cortex, supplementary motor area, or the dorsolateral prefrontal cortex (DLPFC). The stimulation parameters, frequency (Hz),

intensity (mA), and duration, can be tailored to suit specific experimental requirements [33]. In addition, tACS can be delivered across a broad frequency range, typically aligning with standard EEG bands (0.1–80 Hz). The cognitive effects of stimulation vary depending on the frequency used; for example, theta-frequency tACS has been shown to support improvements in working memory, declarative memory, and executive functioning [34]. Alpha-frequency tACS has been shown to improve sustained attention, although it may reduce selective attention in both visual and auditory domains [35]. Gamma-frequency tACS has been found to enhance cognition, working memory, selective attention, motor performance, and learning, with no significant effect on executive functions [36, 37]. As abnormalities in gamma oscillations have been identified as a contributing factor to cognitive impairments in AD [38], recent studies involving gamma-frequency stimulation have provided additional evidence supporting this association [39, 40].

One study demonstrated that gamma oscillation dysfunctions in rodents were a direct consequence of the disease pathology [41]. Additional research reported altered gamma activity in both Alzheimer’s patients and mouse models of the disease [42, 43], with other studies similarly observing reductions in gamma-band activity [44]. In response to these findings, one investigation employed brain stimulation at three different frequencies, 20, 40, and 80 Hz, to boost gamma oscillations and improve cognitive function in mice, revealing that 40 Hz stimulation led to significant improvements [45]. Another study supported these results, showing that 40 Hz stimulation, compared to 8 Hz and 80 Hz, produced notable enhancements in cognitive performance [46]. These findings have led to the application of gamma-band stimulation, specifically at 40 Hz, in patients with AD, with evidence supporting its efficacy in enhancing cognitive function [21]. Consequently, 40 Hz tACS has emerged as the most utilized frequency within the gamma range and was employed in the present study to induce gamma oscillations in the brain. Furthermore, previous studies have demonstrated that applying gamma-frequency tACS to the frontal lobes results in improvements in working memory and problem-solving abilities [25]. This is particularly relevant, as work-

ing memory is a fundamental cognitive function frequently compromised in individuals with dementia [47, 21].

1.3.3 Cognitive Exercises (CE)

Cognitive exercise, frequently combined with non-invasive brain stimulation techniques, is regarded as an effective approach to enhance cognitive function and slow the progression of dementia. It involves repeated engagement in tasks that target specific cognitive domains, such as memory, to support and improve mental performance in individuals with dementia. These exercises utilize structured activities specifically designed to reduce the impact of impairments in targeted cognitive abilities [48].

CE can be delivered through various formats, including paper-and-pencil tasks, computer-based programs, or facilitator-guided activities. The application of cognitive exercise principles began in the early 1980s, initially targeting cognitively healthy older adults with subjective cognitive concerns [49, 50, 51, 52, 53]. Evidence suggests that consistent, structured cognitive training programs can help maintain or even improve cognitive abilities in affected individuals [21, 48]. For instance, one study demonstrated that cognitive exercises led to improvements in memory among patients with mild cognitive impairment [54]. However, the benefits of such training appear to be limited primarily to the specific tasks practiced, with minimal transfer to untrained cognitive [55, 50]. As a result, several studies have proposed that combining gamma-frequency tACS with cognitive exercises may yield more substantial and generalized cognitive improvements [21, 56]. Cognitive exercises enhance synaptic strength among neurons, while electrical stimulation further amplifies this effect. Together, these mechanisms contribute to the expansion of cortical representation within the activated neural networks, increased excitability in adjacent brain regions, and potential transfer of improvements to non-trained cognitive tasks [57]. Accordingly, the clinical trial from which the data for this thesis were derived aimed to investigate the combined effects of real and

sham tACS administered simultaneously with CE in patients diagnosed with AD or mixed Alzheimer’s pathology.

1.3.4 Single and multi-Sessions Gamma-tACS Studies

Recent studies have explored the efficacy of single-session gamma-frequency tACS in individuals with mild cognitive impairment (MCI) and Alzheimer’s disease (AD), with a focus on its cognitive and electrophysiological effects. These investigations predominantly employed randomized, sham-controlled designs. One such study implemented a crossover sham-controlled protocol involving 25 participants with MCI and 35 with AD, aiming to assess whether gamma-tACS could serve as a diagnostic marker to distinguish between the two groups [58]. Stimulation was applied to six cortical sites in the left hemisphere across separate sessions, and outcomes were evaluated through resting-state EEG and cognitive assessments. Results indicated significant cognitive improvements in the MCI group, particularly following stimulation of the dorsolateral and dorsomedial prefrontal cortices, whereas no measurable effects were observed in the AD group.

In another double-blind, sham-controlled repeated-measures trial, 20 individuals with MCI received 40 Hz tACS and tDCS targeting the dorsolateral prefrontal cortex for 30 minutes [14]. EEG analyses revealed that tACS induced a notable increase in beta-band activity in the right frontoparietal regions, an effect not observed with either tDCS or sham stimulation.

A third investigation utilized a randomized, double-blind, sham-controlled crossover design to evaluate memory outcomes in 20 patients with mild cognitive impairment due to Alzheimer’s disease (MCI-AD) [59]. Each participant received both active and sham stimulation in separate sessions. Active stimulation consisted of a single 60-minute session of 40 Hz gamma-frequency transcranial alternating current stimulation (tACS) delivered at 3.0 mA peak-to-peak (1.5 mA peak-to-baseline), with the active electrode positioned over Pz,

targeting the medial parietal cortex and precuneus, and an extracephalic reference electrode placed over the right deltoid muscle. Episodic memory was assessed before and after stimulation using the Rey Auditory Verbal Learning (RAVL) test, while associative memory was evaluated during stimulation using a face–name association task (FNAT). In addition, short-latency afferent inhibition (SAI) measured via transcranial magnetic stimulation (TMS) was used as an indirect marker of cholinergic transmission.

The results demonstrated a significant improvement in RAVL total recall and delayed recall scores following gamma-tACS compared with sham stimulation. Associative memory performance during stimulation also improved, and these cognitive enhancements were accompanied by restoration of SAI measures, suggesting modulation of cholinergic circuitry. These findings indicate that a single session of 40 Hz tACS over the medial parietal cortex may transiently enhance episodic memory in individuals with MCI-AD. However, given the pilot nature of the study and the short-term assessment window, further research is required to determine the durability and broader clinical relevance of these effects.

1.4 TACS Crossover Trial Study Design and Participants

This thesis research is part of a study [60, 61] (ClinicalTrial ID: NCT02908815) initiated in January 2021 to investigate the efficacy of transcranial alternating current stimulation (tACS) administered simultaneously with cognitive exercises (CE). Functional near-infrared spectroscopy (fNIRS) was used as a secondary outcome measure in the clinical trial, while the Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog) served as the primary outcome measure. In this section, the protocol of the tACS+CE efficacy study is briefly outlined. The study followed a randomized, double-blind, placebo-controlled crossover design.(Figure 1.2)

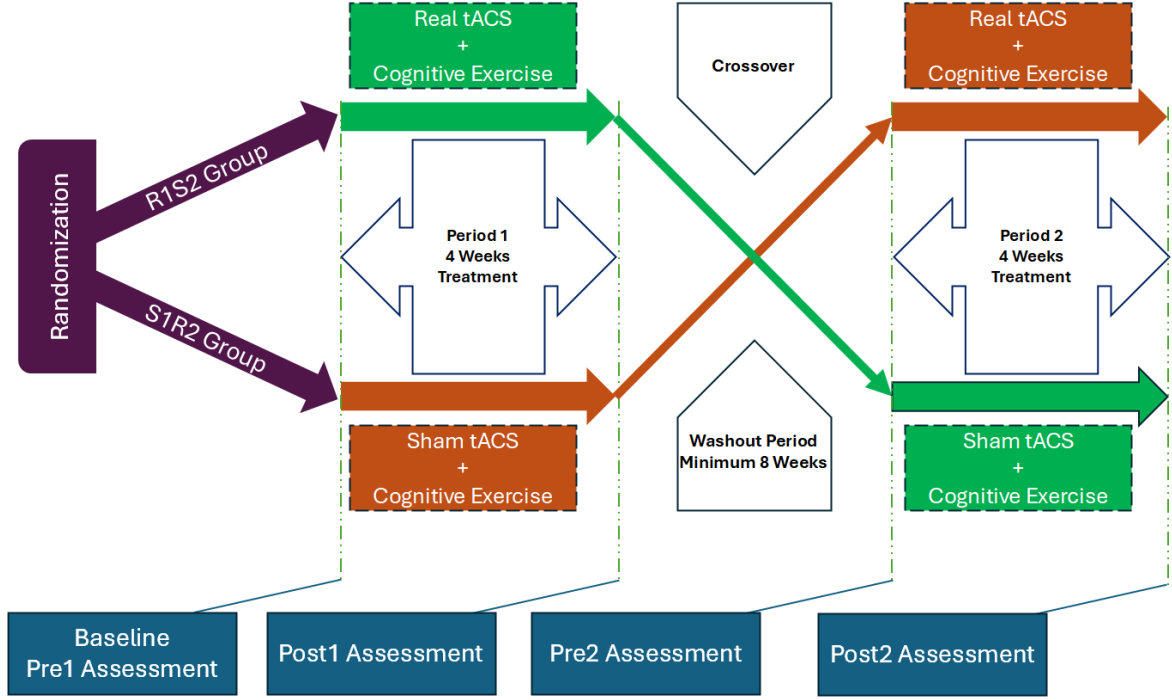


Figure 1.2: Illustration of the crossover study framework used in this thesis

Each participant underwent two treatment phases, each lasting four weeks, during which they received either active or sham tACS concurrently with cognitive exercises (CE). Sessions were conducted five days per week (excluding weekends), with a washout period of 2 to 5 months between the two treatment phases to minimize potential carryover effects. Each treatment day consisted of two 30-minute in-person cognitive exercise sessions administered alongside either real or sham tACS, separated by a 30-minute intersession break. Participants were randomly assigned to either the real/sham or sham/real sequence using a stratified randomization approach, accounting for age and baseline cognitive performance as determined by the Montreal Cognitive Assessment (MoCA) [62]. Table 1.1 presents the descriptive statistics of participants' demographic characteristics stratified by group.

Table 1.1: Descriptive statistics summarizing the demographic characteristics of study participants stratified by group.

Participant Characteristic	Real/Sham Group ($n = 19$)	Sham/Real Group ($n = 22$)	p -Value
Age (years)	73.1 ± 7.6	75.2 ± 9.5	0.44
MoCA	16.6 ± 5.6	15.1 ± 5.3	0.39
MADRS	3.1 ± 3.7	4.5 ± 5.2	0.32
Sex			0.32
Female	8 (42.1%)	6 (27.3%)	
Male	11 (57.9%)	16 (72.7%)	
Handedness			0.46
Right	18 (94.7%)	21 (95.5%)	
Left	1 (5.3%)	0 (0.0%)	
Ambidextrous		1 (4.5%)	
Education			0.51
High school or below	7 (36.8%)	6 (27.3%)	
Above a high school diploma	12 (63.2%)	16 (72.7%)	

Note: Continuous variables are presented as mean $\pm$ standard deviation and were compared using independent samples t -tests. Categorical variables are presented as frequency (percentage) and were compared using chi-square tests or Fisher’s exact tests when expected cell counts were less than five.

The tACS+CE study was reviewed and approved by the Biomedical Research Ethics Board at the University of Manitoba, in accordance with established ethical standards for research involving human participants. A total of 41 participants were included in the analysis for this thesis. Participants were eligible for inclusion if they met the following criteria:

- Aged between 50 and 95 years
- MoCA score between 5 and 25
- Being on a stable dose of dementia-related medication or no medication 3 months prior to the study

- Demonstrated fluency in reading and writing English

Exclusion criteria were designed to minimize confounding factors and ensure participant safety. Individuals were excluded if they had:

- A diagnosis of Parkinson’s disease, Parkinsonian dementia, Huntington’s disease, significant aphasia, intellectual disability, major depressive or anxiety disorder, bipolar disorder, schizophrenia, or any other severe psychiatric condition
- A history of epilepsy or epileptic seizures
- Difficulty communicating effectively in English
- Severe visual or auditory impairments that could interfere with cognitive assessments
- A current diagnosis of a substance use disorder
- Concurrent participation in another therapeutic trial related to dementia
- Plans to initiate or modify dementia-related medication during the study periods

The transcranial alternating current stimulation (tACS) intervention was delivered using the Soterix Medical 1×1 tACS device. Figure 1.3 illustrates the tACS device and the head strap used to secure the electrodes. The image is reproduced from the manufacturer’s website with attribution. Source: Soterix Medical Inc. [63].

The stimulation protocol implemented in this study included the following parameters:

- Stimulation frequency: 40 Hz
- Session duration: 30 minutes
- Current intensity: 0.75 mA (equivalent to 1.5 mA peak-to-peak)

- Electrode montage: Bipolar configuration with the anode positioned at F3 (targeting the left dorsolateral prefrontal cortex) and the cathode at Fp2 (right supraorbital region), based on the international 10–20 EEG system
- Waveform: Sinusoidal
- Sham condition: Simulated using a brief on/off ramp to replicate the sensation of stimulation without delivering sustained current

For the sham condition, the tACS device (Soterix Medical Inc.) was programmed to simulate stimulation without delivering continuous current. Current intensity was gradually increased for 30 seconds and then reduced to zero over the following 30 seconds, after which no current was delivered for most of the session. To maintain blinding, a brief ramp-up and ramp-down sequence was repeated during the final minute of stimulation.

Saline-moistened sponge electrodes were placed over the left dorsolateral prefrontal cortex (F3 according to the international 10–20 EEG system) and the contralateral supraorbital region. Study personnel were trained to confirm adequate electrode contact by monitoring device feedback throughout each session. All participant-reported sensations and observed adverse effects, including discomfort, skin irritation, burning sensations, fatigue, headache, dizziness, disorientation, tingling, pain, or other unexpected symptoms, were systematically recorded during stimulation.

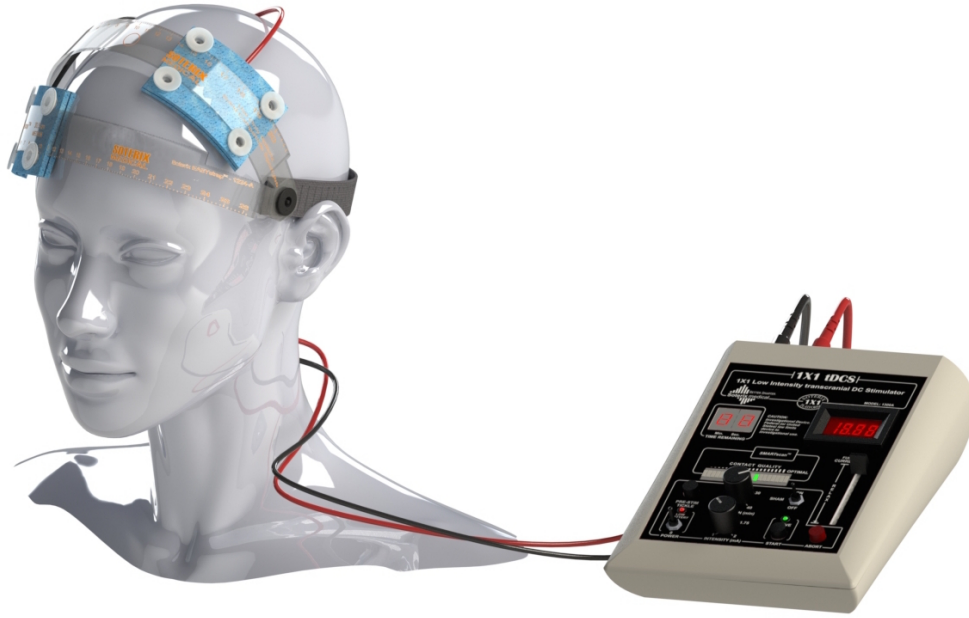


Figure 1.3: Soterix Medical 1×1 transcranial alternating current stimulation (tACS) device and EASYstrap head strap used for electrode placement. (Source: Soterix Medical Inc., reproduced for educational purposes; permission not required.)

Participants were instructed and helped by the treatment administrators to perform cognitive exercises using the MindTriggers app while receiving the real or sham tACS on an iPad. The cognitive exercises (CE) administered during each treatment session targeted multiple cognitive domains, including spatial cognition, associative memory, visual memory, image–word association, categorization, and sentence memorization and recall. These exercises were delivered using the MindTriggers application, developed by Z. Moussavi [64], which served as the cognitive training platform in this study. The application comprises seven distinct tasks designed to stimulate different aspects of cognitive functioning, such as associative and visual memory, spatial memory, and interhemispheric connectivity.

During each treatment session, participants completed three of the seven available exercises. All sessions were supervised by a trained tutor who remained present throughout to provide guidance and ensure consistent task administration, following the study protocol described in [21]. Although the MindTriggers application is available across multiple plat-

forms, all participants used 13-inch iPads to maintain consistency and standardization across sessions.

Overall, MindTriggers provides a structured set of experimental tasks aimed at supporting and strengthening working memory and related cognitive functions. Spatial memory tasks assist individuals in navigating and orienting themselves within unfamiliar environments, while associative memory exercises focus on forming links between concepts or objects encountered in daily life, such as associating an automated teller machine with money. With aging and in individuals with dementia, the ability to establish new associations or learn unfamiliar concepts often declines. Verbal fluency, defined as the capacity to retrieve and articulate words, is similarly affected in dementia populations. Each game within the application includes multiple difficulty levels, some of which are intentionally challenging even for cognitively healthy individuals, allowing the exercises to be adapted to varying cognitive abilities. Although the platform is specifically designed for individuals with dementia, it incorporates scalable complexity to accommodate individual differences. Detailed guidelines for administering these exercises to participants with dementia are provided in the study protocol and summarized in [21].

These parameters were selected to align with previous research supporting the use of gamma-frequency tACS in cognitive intervention protocols for dementia. Individuals were recruited either from earlier studies or through recommendations from family doctors, geriatric specialists, neurologists, and psychiatrists, who also supplied the diagnostic reports. Prior to taking part in the study, all participants (along with a family member or legal guardian) provided written informed consent, as approved by the Biomedical Research Ethics Board at the University of Manitoba. Additionally, before the initial assessment, participants were evaluated for their capacity to tolerate tACS. If a participant could not tolerate tACS, they were given the alternative to engage solely in cognitive exercises; however, their performance data were excluded from the analysis in this paper. Finally, every participant was assigned a unique anonymized code for data recording to maintain confidentiality. The study co-

ordinator handled both the enrollment and group assignment of the participants using a computerized randomization system.

Participants underwent evaluations 3 to 7 days prior to the first session and again 3 to 7 days following the final session of each treatment block. These assessments were conducted by an evaluator who was blinded to the type of stimulation administered (real or sham). The primary outcome measure was the change in scores on the Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog), a well-established instrument for evaluating cognitive function in individuals with Alzheimer’s disease or other forms of dementia. The ADAS-Cog includes 11 tasks designed to assess various cognitive abilities such as memory, language, orientation, and praxis, providing a comprehensive overview of cognitive status [65]. Scores on the ADAS-Cog range from 0 to 70, where lower scores denote better cognitive function, and higher scores indicate more severe impairment.

1.5 Research Objectives

The objectives of this thesis are as follows:

1. To evaluate the treatment effects of transcranial alternating current stimulation (tACS) combined with cognitive exercises using a randomized, double-blind 2×2 crossover design.
2. To assess treatment, period, and carryover effects using classical frequentist methods, including parametric t -tests, ANOVA, and ANCOVA models.
3. To examine the robustness of classical results using bootstrap resampling techniques.
4. To develop and apply a Bayesian framework for crossover trials, providing probabilistic inference for treatment effects.

1.6 Thesis Organization

This thesis begins with an overview of crossover study designs, highlighting their statistical efficiency, advantages for within-subject comparisons, and common methodological challenges such as carryover effects and assumption violations. The motivation for employing multiple analytical frameworks to evaluate treatment effects in a 2×2 crossover trial is also introduced.

Chapter 2 examines classical statistical methods, focusing on parametric approaches commonly used to estimate treatment, period, and carryover effects. Model assumptions, including normality and homogeneity of variance, are assessed using diagnostic techniques to evaluate model adequacy.

Chapter 3 introduces bootstrap resampling as a nonparametric alternative for statistical inference. The use of bootstrap confidence intervals and hypothesis testing is emphasized, particularly in the context of small sample sizes and potential deviations from standard distributional assumptions.

Chapter 4 presents a Bayesian framework for analyzing crossover data, detailing model specification, prior selection, and posterior inference. Bayesian model comparison and posterior probabilities are used to provide interpretable evidence regarding treatment effects.

Finally, Chapter 5 summarizes the findings across all analytical approaches, compares their strengths and limitations, and discusses methodological implications and directions for future research on crossover trials in dementia studies.

Chapter 2

Frequentist analyses of 2×2 crossover studies

Summary. This chapter presents classical frequentist analyses of the randomized, double-blind, placebo-controlled 2×2 crossover trial comparing real versus sham transcranial alternating current stimulation (tACS), combined with cognitive exercises, in individuals with dementia. Within the general linear model framework, we estimate treatment effects while accounting for period, sequence, and potential carryover effects. Model assumptions are examined using graphical diagnostics and formal tests before applying parametric methods. Paired-sample t-tests assess within-subject differences between real and sham tACS, whereas independent-samples t-tests evaluate baseline equivalence across treatment sequences. Subsequently, two-way ANOVA quantifies the contributions of treatment, period, sequence, and carryover, and ANCOVA models incorporating baseline cognitive scores provide covariate-adjusted treatment effects.

2.1 Introduction

In this chapter, we analyze data from the randomized, double-blind, placebo-controlled 2×2 crossover trial of real versus sham tACS combined with CE in individuals with dementia, as described in Chapter 1. Our main focus is on quantifying treatment effects while explicitly adjusting for period, sequence, and other design-related features of the crossover structure. The resulting classical analyses serve as a methodological benchmark against which the bootstrap and Bayesian procedures developed in subsequent chapters can be compared [1, 2].

Under the frequentist framework, parameters such as means and treatment effects are considered fixed but unknown quantities, and probability is defined as the long-run frequency of events under repeated sampling [66, 67]. Inference is based on the sampling distribution of test statistics, allowing conclusions to be drawn about the population from observed data. Confidence intervals provide an estimated range for the true effect, and hypothesis tests determine whether observed differences are likely to reflect genuine effects or random variation [68]. Within this approach, smaller p -values indicate stronger evidence against the null hypothesis, guiding the decision to accept or reject it at a chosen significance level (commonly $\alpha = 0.05$).

Before applying frequentist analyses, it is essential to verify that the main assumptions underlying these models are met. Therefore, the chapter begins by evaluating normality, homogeneity of variances, and independence of residuals to ensure that the results of subsequent analyses are statistically valid. Normality is assessed both visually, through Q-Q plots and histograms [69], and statistically, using the Shapiro–Wilk, which is known for its high power in small samples [70]. Homogeneity of variances across treatment conditions is tested using Levene’s test [71]. These diagnostic checks help determine whether the data are suitable for classical parametric analyses or whether deviations might influence the robustness of the conclusions [72]. After verifying assumptions, t-tests are conducted as preliminary analyses. The paired-sample t-test examines within-subject differences between

real and sham stimulation conditions, making use of the repeated-measures design of the crossover study. The independent-samples t-test evaluates baseline equivalence between sequence groups (e.g., AB vs. BA), confirming the success of randomization and ensuring that no pre-existing differences exist before treatment [73].

Following these comparisons, a two-way Analysis of Variance (ANOVA) is applied to test for the main effects of treatment, period, and sequence, as well as potential carryover effects from one period to the next [2, 72]. ANOVA partitions the total variation in the data into systematic components associated with experimental factors and random error, allowing assessment of whether mean responses differ significantly across conditions.

Finally, an Analysis of Covariance (ANCOVA) extends this framework by incorporating baseline cognitive scores. Adjusting for covariates reduces unexplained variability and provides a more accurate estimate of treatment effects that accounts for individual differences. This approach yields adjusted means that better represent the net effect of tACS, independent of participants' initial cognitive levels or other sources of variability [74, 75].

All analyses presented in this chapter are carried out within the general linear model framework, with diagnostic checks performed to verify that model assumptions are satisfied. The results obtained here form the foundation for the resampling-based and Bayesian analyses presented in the next chapters, which address some of the limitations of classical methods, particularly their reliance on distributional assumptions and sensitivity to small sample sizes.

2.2 Data Preparation and Descriptive Statistics

The dataset consists of repeated ADAS-Cog measurements collected within a 2×2 crossover framework. Participants were randomly assigned to one of two sequences: Sequence AB, who received real tACS in the first period followed by sham stimulation in the second, and

Sequence BA, who received sham stimulation first and real stimulation in the second period. Each participant contributed four ADAS-Cog observations: baseline and post-treatment scores for Period 1, and baseline and post-treatment scores for Period 2. These raw values formed the basis for all subsequent classical analyses.

2.2.1 Dataset Description

A total of $N = 41$ participants completed both periods of the study, with $n_{AB} = 19$ randomized to Sequence AB and $n_{BA} = 22$ to Sequence BA. For each period, ADAS-Cog was assessed immediately prior to stimulation (baseline) and again at the end of the treatment session. Since higher ADAS-Cog values indicate greater cognitive impairment, reductions in scores over time reflect improvement.

2.2.2 Computation of Within-Period Change Scores

To quantify the immediate effect of stimulation within each period, change scores were calculated by subtracting each baseline value from the corresponding post-treatment score. Specifically, for Period $i \in \{1, 2\}$,

$$\Delta_{P_i} = \text{ADAS-Cog}_{\text{post}, P_i} - \text{ADAS-Cog}_{\text{baseline}, P_i}.$$

Negative values of Δ correspond to improvement. Although the crossover structure allows analyses based on either adjusted post-treatment scores or within-period differences, reporting change scores is conventional in crossover trials as they directly reflect short-term treatment response. Moreover, examining Δ reduces between-subject variability that may arise from individual differences in baseline cognitive status, which is particularly relevant in aging and dementia populations.

2.2.3 Treatment and Period Labels

For statistical modelling, all post-treatment scores and change scores were indexed by period and treatment condition. In Sequence AB, the real stimulation indicator was coded in Period 1 and the sham indicator in Period 2; in Sequence BA the opposite pattern applied. This coding structure allowed the data to be reorganized into treatment-specific and period-specific subsets when required, for example, when performing paired comparisons of real versus sham stimulation within subjects, or when assessing period and sequence effects that are intrinsic to crossover designs.

2.2.4 Descriptive Statistics

Descriptive statistics for the ADAS-Cog measurements are summarized in Table 2.1, stratified by sequence.

Table 2.1: Descriptive statistics for ADAS-Cog scores by sequence

Sequence	Baseline_ P_1		Treated_ P_1		Baseline_ P_2		Treated_ P_2		Delta_ P_1		Delta_ P_2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
AB	22.69	9.11	21.28	8.89	20.37	7.84	19.53	7.22	-1.41	3.84	-0.83	3.81
BA	23.76	11.73	22.43	11.85	22.79	10.19	21.23	11.97	-1.33	5.11	-1.56	5.23

In Sequence AB, the mean baseline scores were 22.70 (SD = 9.12) in Period 1 and 20.38 (SD = 7.84) in Period 2. The corresponding post-treatment means were 21.28 (SD = 8.90) in Period 1 and 19.54 (SD = 7.22) in Period 2, respectively. The average within-period change was -1.41 (SD = 3.84) for Δ_{P_1} and -0.84 (SD = 3.82) for Δ_{P_2} , indicating modest average improvement in both periods. Sequence BA showed similar patterns: baseline means were 23.76 (SD = 11.74) in Period 1 and 22.79 (SD = 10.20) in Period 2, with post-treatment means of 22.43 (SD = 11.85) and 21.23 (SD = 11.98), respectively. The mean change scores

were -1.33 (SD = 5.11) for Δ_{P1} and -1.56 (SD = 5.24) for Δ_{P2} . Overall, both sequences demonstrated small average improvements, with more variability observed in Sequence BA. The distributional patterns observed in the boxplots (Figures 2.1 and 2.2) are consistent with these numerical summaries.

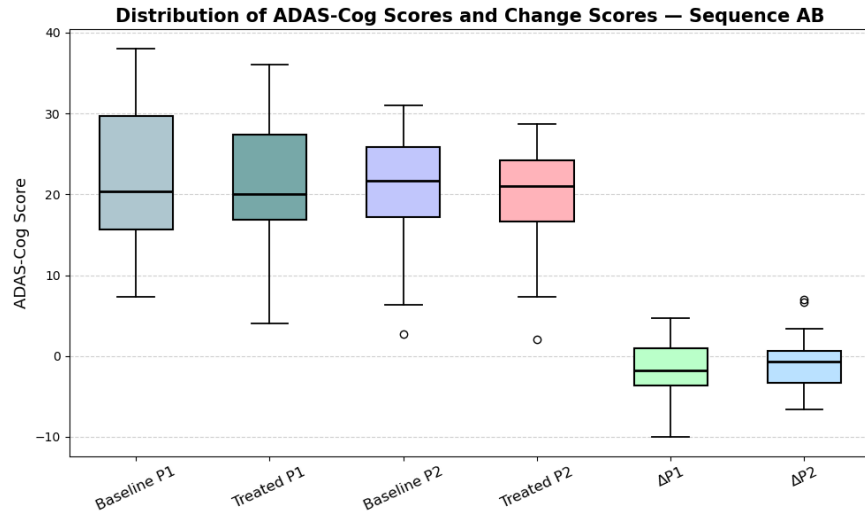


Figure 2.1: Boxplots of ADAS-Cog Scores and Within-Period Change Scores by Sequence AB

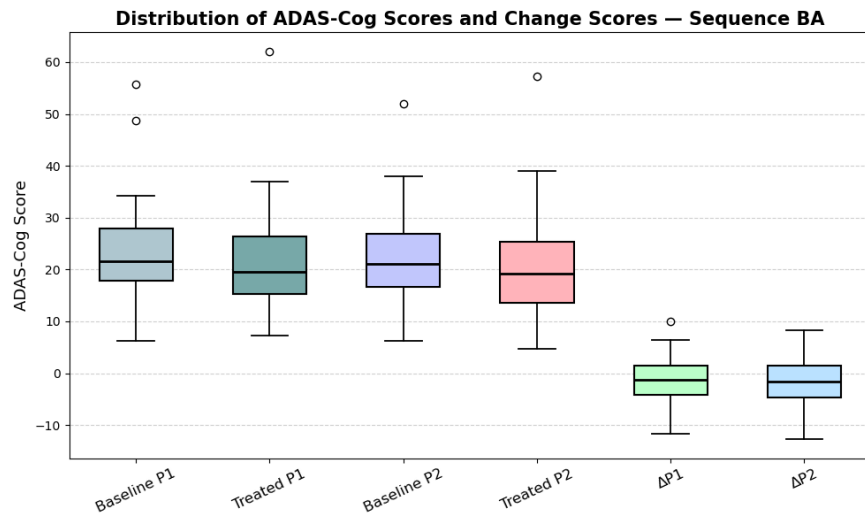


Figure 2.2: Boxplots of ADAS-Cog Scores and Within-Period Change Scores by Sequence BA

In Sequence AB, both baseline and post-treatment scores in each period appear rel-

atively compact, with most observations falling between approximately 10 and 30 and only a few values extending beyond this range. The change scores for AB are centered slightly below zero, with the interquartile range corresponding to modest improvements of roughly 1–3 points.

In contrast, Sequence BA demonstrates noticeably wider score distributions, particularly in Period 1, where values range from single-digit scores to above 50. The associated change-score boxplots for BA show a substantially broader interquartile range and several extreme values in both negative and positive directions, indicating greater individual variability within this sequence. The histograms of the change scores depicted in Figure 2.3 provide additional context.

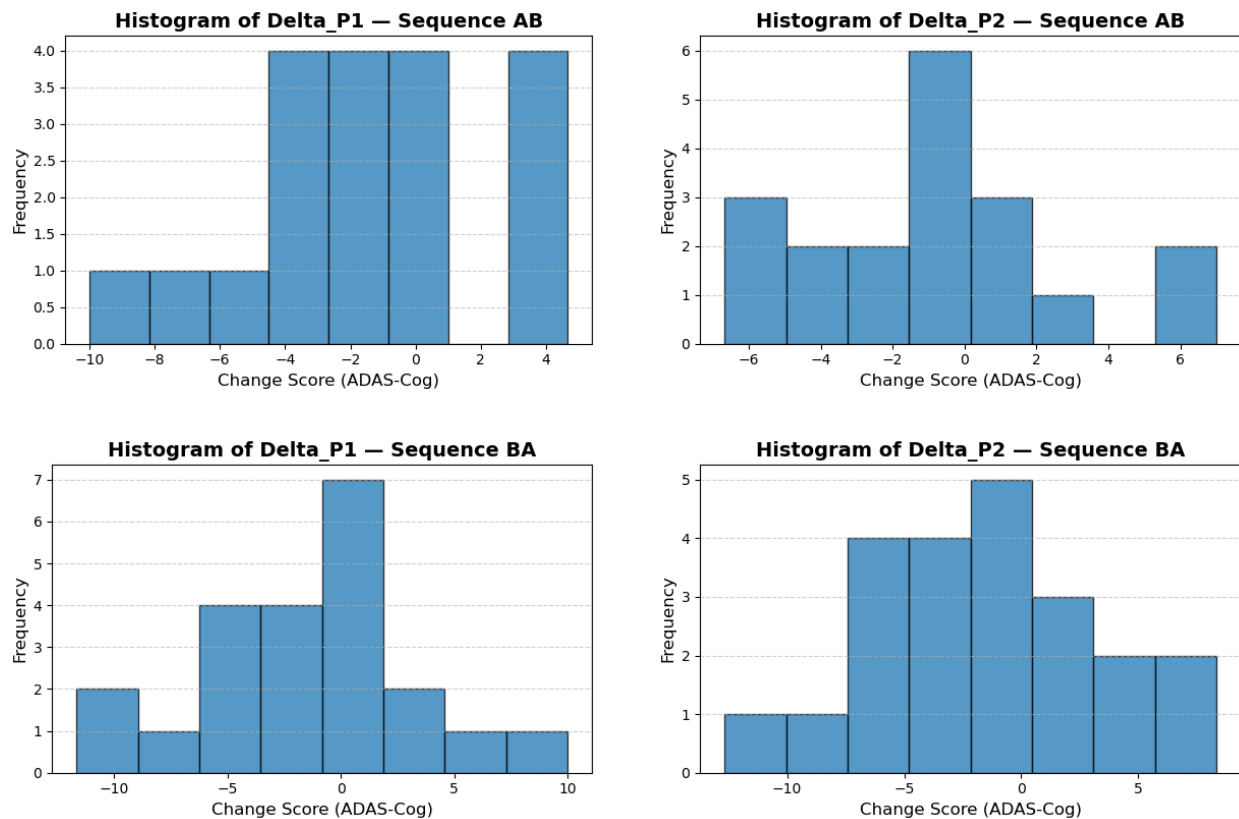


Figure 2.3: Histograms of ADAS-Cog change scores across sequences (AB and BA).

Sequence AB shows relatively tight, unimodal distributions, with most Δ_{P1} and Δ_{P2} values clustered close to zero and only a few pronounced improvements or declines. In

contrast, Sequence BA exhibits noticeably wider spreads in both periods, with Δ values ranging from substantial improvements (approximately -10 to -12) to moderate worsening ($+8$ to $+10$). Despite this broader range, the distributions remain roughly centered near zero, indicating that the increased variability reflects greater heterogeneity among participants rather than systematic shifts.

2.2.5 Initial Observations from the Data

Examination of the descriptive patterns indicates that the two sequences began the study with broadly comparable baseline ADAS-Cog levels, supporting the adequacy of the randomization. Across both periods, average change scores were modest and negative, suggesting slight overall improvement irrespective of stimulation order. This general trend, observed in both sequences, may reflect non-specific influences such as repeated testing, familiarity with study procedures, or short-term cognitive fluctuations.

More noteworthy is the difference in variability between sequences. Sequence AB exhibited relatively consistent responses, with change scores concentrated within a narrow range. In contrast, Sequence BA showed appreciably wider dispersion, including participants with substantial improvements as well as others with moderate worsening. This broader spread indicates greater heterogeneity in response profiles rather than systematic differences in mean performance.

Taken together, the initial descriptive findings point to balanced baseline characteristics, small average improvements across periods, and higher variability in Sequence BA, features that should be considered when evaluating treatment, period, and sequence effects in the subsequent inferential analyses.

2.3 Assumption Checking

The classical parametric methods used in this chapter, paired and independent t-tests, ANOVA, and ANCOVA, rely on several underlying assumptions. In particular, these methods assume that the residuals are approximately normally distributed, that variances are homogeneous across groups where appropriate, and that observations are independent. Violation of these assumptions can lead to biased tests, invalid p -values, or reduced power, which is especially important in small-sample clinical studies such as the present 2×2 crossover trial.

In this study, ADAS-Cog scores were obtained at baseline and at the end of each treatment period. Participants were randomized to one of two sequences: Group A received real tACS in Period 1 and sham tACS in Period 2, whereas Group B received sham tACS in Period 1 and real tACS in Period 2. For the classical analyses, the primary outcomes considered in this chapter are the post-period ADAS-Cog scores and, where relevant, the within-subject changes from baseline to post-treatment within each period. Assumption checks were therefore performed on (i) the distributions of ADAS-Cog outcomes within each sequence and period, and (ii) the residuals from the models fitted in subsequent sections.

2.3.1 Assessment of Normality

Normality was assessed using a combination of formal statistical tests and visual inspection. The Shapiro–Wilk test was used to evaluate whether the sample distributions deviated significantly from a theoretical normal distribution. This test is recommended for small to moderate sample sizes and is widely applied in biomedical research to assess normality of continuous outcomes [76]. A non-significant p -value ($p\text{-value} > 0.05$) indicates insufficient evidence to reject normality, whereas a significant p -value suggests potential skewness, heavy tails, or other departures from a Gaussian distribution.

To complement the formal test, visual diagnostics were performed using Q–Q (quantile–quantile) plots of the baseline ADAS-Cog scores (Figure 2.4). In these plots, observations that align closely with the 45-degree reference line indicate approximate normality, while systematic deviations, such as curvature or departures in the tails, suggest skewness or kurtosis issues [77].

For the AB sequence, the Shapiro–Wilk test showed no evidence of deviation from normality ($W = 0.953$, $p\text{-value} = 0.445$). The corresponding Q–Q plot demonstrated points closely following the reference line, with only minor deviations in the tails, which is typical in small samples. Together, these findings indicate that the baseline ADAS-Cog scores in the AB group are approximately normally distributed.

In contrast, for the BA sequence, the Shapiro–Wilk test indicated a statistically significant deviation from normality ($W = 0.893$, $p\text{-value} = 0.022$). The Q–Q plot showed noticeable curvature, particularly in the upper tail, where several participants had elevated baseline ADAS-Cog scores. This pattern is consistent with mild right-skewness and heavier-than-normal upper-tail behavior.

Overall, the normality assumption holds for the AB sequence but is not satisfied for the BA sequence. Nevertheless, t-tests are generally robust to moderate violations of normality when group sizes across groups are comparable. Therefore, the baseline comparison remains appropriate. Importantly, the non-normality observed in the BA sequence highlights a limitation of classical parametric approaches and further motivates the use of bootstrap techniques in Chapter 3, which do not rely on distributional assumptions.

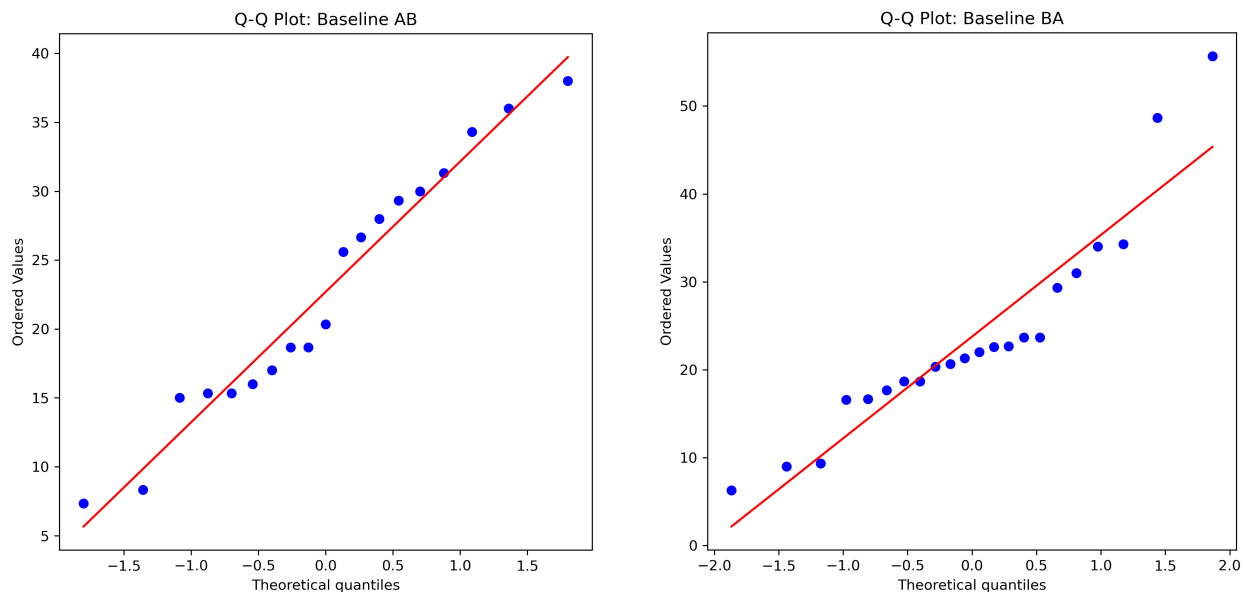


Figure 2.4: Q-Q plots of baseline ADAS-Cog scores for Sequence AB (left) and Sequence BA (right).

2.3.2 Homogeneity of Variances

Before comparing baseline ADAS-Cog scores between the two treatment sequences, it is important to assess whether the groups exhibit similar variability. Many classical statistical procedures, including the independent samples t-test and the standard two-sample components of ANOVA, rely on the assumption that the groups being compared have approximately equal variances (homoscedasticity). Violation of this assumption can lead to biased standard errors and incorrect inference, particularly when sample sizes differ between groups [78].

To evaluate homogeneity of variances between the AB and BA sequences at Period 1 baseline, Levene’s test was applied. Levene’s procedure tests the null hypothesis that the variances of the two groups are equal. A non-significant p -value (p -value > 0.05) indicates no evidence of heterogeneous variances, whereas a significant result suggests that the groups differ in variability and that methods robust to heteroscedasticity should be used. Because one of the groups (BA) showed evidence of non-normality (Section 2.3.1), the median-based

version of Levene’s test was used, as it provides greater robustness under skewed distributions and outliers [79].

The test indicated no evidence of unequal variances between groups (Levene’s $W = 0.0008$, $p\text{-value} = 0.9769$). This extremely high p -value suggests that the spread of baseline ADAS-Cog scores is nearly identical in the two sequences. Consequently, variability differences are unlikely to influence the subsequent baseline comparison in Section 3.3, and the homogeneity of variances assumption can be considered satisfied at this stage of the analysis.

2.3.3 Baseline Equality Between Sequences

An essential assumption in the analysis of two-period, two-treatment crossover trials is that the randomized sequences begin the study at comparable baseline levels. If one sequence enters the trial with systematically higher or lower cognitive impairment, subsequent comparisons of treatment effects may be biased, even when randomization has been applied. Baseline imbalance can distort treatment estimates, confound period effects, and obscure the detection of potential carryover effects [2]. Therefore, verifying baseline equivalence is a crucial step before any classical treatment comparisons are performed.

To evaluate whether the AB and BA sequences were comparable at the start of Period 1, baseline ADAS-Cog scores were compared between the two groups. Because earlier analyzes (Section 2.2.1) indicated a violation of normality in the BA sequence and given the modest sample sizes ($n_{AB} = 19$ vs. $n_{BA} = 22$), a nonparametric approach was preferred. The Mann–Whitney U test, also known as the Wilcoxon rank-sum test, was selected, as it does not rely on distributional assumptions and is appropriate for comparing continuous outcomes when the assumption of normality is questionable [80].

The Mann–Whitney U test evaluates whether the distributions of baseline ADAS-Cog scores differ between sequences. A non-significant result suggests that both sequences

begin the study with comparable levels of cognitive impairment, supporting the validity of subsequent crossover analyses. A significant result would indicate baseline imbalance, requiring caution in interpreting classical treatment effects.

The Mann–Whitney U test indicated no evidence of a baseline difference between the two sequences ($U = 200.5$, $p\text{-value} = 0.8342$). This large p -value suggests that the distributions of baseline ADAS-Cog scores in the AB and BA groups are comparable, and any observed differences at the start of the first period are likely attributable to random variation rather than systematic imbalance. As a result, the assumption of baseline equivalence is satisfied, supporting the validity of subsequent classical analyses of treatment, period, and sequence effects.

2.4 Testing for Design-Specific Effects

Classical analysis of a 2×2 crossover design requires assessing several design-specific effects, such as, sequence, period, and carryover effects, before drawing conclusions about treatment efficacy. These checks ensure that treatment comparisons are not biased by systematic differences in participant allocation, time-related changes, or lingering effects of the prior treatment. The descriptive statistics in Table 1 provide the basis for evaluating these properties.

2.4.1 Sequence Effect

A sequence effect evaluates whether participants differed systematically between the two treatment orders (AB vs. BA) at the beginning of the study, before any carryover could occur. In a 2×2 crossover design, this assessment is based solely on Period 1 outcomes. Because baseline differences can obscure early treatment comparisons, the analysis here uses the Period-1 change scores (Δ_{P1}), defined as the difference between post-treatment

and baseline ADAS-Cog values. This approach provides a direct measure of the immediate response in Period 1 and adjusts for individual variation in initial cognitive status.

An independent samples t -test was conducted to compare Δ_{P1} between the two sequences. The mean change in Period 1 was -1.41 ($SD = 3.84$) for Sequence AB and -1.33 ($SD = 5.11$) for Sequence BA. Statistical comparison of these values yielded $t(39) = -0.059$, p -values = 0.953 , indicating no evidence of a sequence effect.

2.4.2 Period Effect

A period effect refers to any systematic difference in participant performance between Period 1 and Period 2 that is unrelated to the treatment itself. In a crossover design, such effects may arise from factors such as increased familiarity with testing procedures, practice effects, fatigue, or broader temporal influences. To examine this possibility, each participant's change score in Period 1 (Δ_{P1}) was compared with their corresponding change score in Period 2 (Δ_{P2}), allowing a direct assessment of whether cognitive responses shifted simply as a function of time.

Across the sample, the average change in Period 1 was -1.37 ($SD = 4.51$), while the mean change in Period 2 was -1.23 ($SD = 4.59$), indicating very similar overall patterns of improvement. A paired t -test comparing the two sets of change scores confirmed that this difference was not statistically meaningful, $t(40) \approx -0.16$, p -value ≈ 0.87 .

Overall, there was no indication of a period effect. Participants responded similarly across the two periods, suggesting that temporal factors are unlikely to influence the interpretation of the treatment comparisons that follow.

2.4.3 Carryover Effect

The classical assessment of carryover in a 2×2 crossover design follows the procedure introduced by Grizzle (1965), which evaluates whether the treatment administered in Period 1 has a residual influence on outcomes measured in Period [6]. Under this method, a carryover effect is tested by comparing the Period 2 change scores (Δ_{P2}) between the two treatment sequences. Since Sequence AB receives real stimulation in Period 1 and sham in Period 2, while Sequence BA receives the opposite order, any lingering influence of the first treatment would be reflected as a systematic difference in Period 2 responses between the sequences.

In this dataset, the mean Period 2 change for Sequence AB was -0.84 ($SD = 3.82$), whereas Sequence BA showed a mean of -1.56 ($SD = 5.24$). Although these values differ slightly in magnitude, Levene’s test indicated no evidence of unequal variances between the sequences ($p\text{-values} = 0.9769$), supporting the use of the standard pooled two-sample t -test. The comparison yielded $t(39) \approx 0.51$, $p\text{-values} \approx 0.61$, indicating that the Period 2 responses did not differ significantly between the sequences.

In summary, the Grizzle’s carryover test provides no evidence of residual treatment effects from Period 1. This supports the assumption that Period 2 outcomes are not biased by prior treatment exposure, reinforcing the validity of the crossover framework for estimating the treatment effect.

2.5 Treatment Effect Analysis

Having verified that no meaningful sequence, period, or carryover effects were present, the data can be analyzed under the standard assumptions of the 2×2 crossover framework. This section focuses on estimating the treatment effect of real tACS relative to sham stimulation using classical frequentist methods. Because each participant received both treatments across two periods, the crossover design allows treatment comparisons to be made within

individuals, providing increased statistical efficiency and reducing between-subject variability. To capture the immediate cognitive response to each stimulation session, all analyses are conducted using within-period change scores ($\Delta = \text{Treated} - \text{Baseline}$), which directly quantify improvement or worsening relative to each participant’s starting point.

The treatment effect is evaluated using several complementary approaches, including a paired t -test comparing within-subject responses to real versus sham stimulation, a between-sequence comparison of treatment differences, and classical crossover analyses based on ANOVA and ANCOVA models. Together, these analyses provide a comprehensive evaluation of the short-term impact of tACS on ADAS-Cog performance.

2.5.1 Paired t -test for Within-Subject Treatment Effect

To examine the short-term effect of real tACS relative to sham stimulation within each treatment sequence, paired t -tests were conducted separately for Sequences AB and BA. In both cases, the outcome measure was the within-period change score $\Delta = \text{post-treatment} - \text{baseline}$, with negative values indicating improvement. For Sequence AB, real stimulation was delivered in Period 1 and sham in Period 2, so the Period 1 change scores represent the response to real tACS and the Period 2 change scores represent the response to sham. In Sequence BA, this pattern was reversed, with sham administered in Period 1 and real stimulation in Period 2.

In Sequence AB, the mean change following real tACS was -1.41 , compared with -0.84 following sham stimulation. A paired t -test comparing these responses showed that the difference was not statistically significant, $t(18) = -0.47$, $p\text{-value} = 0.646$. Although the numerical trend favored real stimulation, the variability of responses within the sequence was large enough that the observed difference could plausibly be attributed to random fluctuation.

A similar pattern emerged in Sequence BA, where the order of administration was

reversed. The mean change following sham stimulation in Period 1 was -1.33 , whereas the change following real stimulation in Period 2 was -1.56 . Again, the paired t -test indicated no significant difference between the two conditions, $t(21) = -0.18$, $p\text{-value} = 0.859$, suggesting that responses to real and sham stimulation were comparable within this sequence as well.

Taken together, the within-sequence analyses do not provide evidence of a differential short-term treatment effect of real tACS compared with sham, despite small numerical advantages in favor of real stimulation in both sequences.

2.5.2 Treatment Comparison with Two-Sample t -test

Although the crossover design allows the primary treatment effect to be evaluated within subjects, it is also informative to examine whether the magnitude of the treatment response differed between the two sequences. This provides an additional check on whether the order of receiving real and sham stimulation influenced participants' overall responses. For each participant, a treatment-difference score was calculated as

$$D = \Delta_{\text{Real}} - \Delta_{\text{Sham}},$$

where negative values indicate greater improvement following real tACS. In Sequence AB, this score was computed as

$$D_{\text{AB}} = \Delta_{P1} - \Delta_{P2},$$

since real stimulation was administered in Period 1. In Sequence BA, the order was reversed, and the treatment difference was

$$D_{\text{BA}} = \Delta_{P2} - \Delta_{P1}.$$

The mean treatment difference was -0.575 in Sequence AB and -0.230 in Sequence BA, indicating a small numerical advantage for real stimulation in both groups. Levene’s test showed no evidence of unequal variances between sequences ($W = 0.739$, $p\text{-value} = 0.395$), supporting the use of a pooled two-sample t -test. Comparison of the treatment-difference scores revealed no significant difference between the two sequences,

$$t(39) = -0.193, \quad p\text{-value} = 0.848.$$

Overall, the between-sequence analysis indicated that the magnitude of the treatment response was similar in the two groups, and that treatment order did not materially influence the estimated effect of real versus sham tACS.

2.5.3 Repeated Measures ANOVA

To assess the effect of real versus sham tACS within the crossover structure, the data were analyzed using the classical 2×2 crossover ANOVA originally described by [6]. This type of ANOVA is specifically designed for crossover trials, where participants receive treatments in different orders and contribute repeated measurements across periods. Unlike standard repeated-measures ANOVA, the classical crossover formulation accounts explicitly for the sequence in which treatments are received and for subjects nested within sequence, which provides the correct error term for testing treatment effects.

The logic of ANOVA rests on partitioning the total variability in the data into meaningful components. The central idea is that differences in group means must be evaluated relative to the variability within the groups themselves; ANOVA therefore compares between-group variance to within-group variance [81]. This variance-based framework is particularly beneficial in multifactor experiments, such as crossover trials, where treatment, period, sequence, and subject-specific effects all contribute to the observed data [82]. In this study,

the response variable was the within-period change score,

$$\Delta = \text{post-treatment} - \text{baseline}$$

representing each participant's short-term cognitive change following real or sham stimulation. Consistent with the classical crossover framework, the following linear model was fitted:

$$\Delta_{ijk} = \mu + \text{Sequence}_i + \text{Subject (Sequence)}_{ij} + \text{Period}_k + \text{Treatment}_{ik} + \varepsilon_{ijk},$$

where:

- μ is the overall mean change,
- Sequence_i (AB or BA) is a between-subject factor,
- $\text{Subject(Sequence)}_{ij}$ is the subject effect nested within sequence (the correct error term for the treatment effect),
- Period_k represents Period 1 vs. Period 2,
- Treatment_{ik} denotes real vs. sham stimulation, and
- ε_{ijk} is the residual error.

This model partitions the total variability into contributions from sequence, period, treatment, and individual subjects, allowing the treatment effect to be tested after accounting for these sources of variation.

The ANOVA results indicated that period did not influence the magnitude of change, $F(1, 39) = 0.037$, $p\text{-value} = 0.848$, showing that responses were similar across Period 1 and Period 2. The treatment effect was likewise not significant, $F(1, 39) = 0.203$, $p\text{-value}$

= 0.655. Although real stimulation produced slightly more negative mean change scores (representing small improvements), this difference was small relative to the subject-level variability. The subject-within-sequence component was large relative to the residual error, $F(80, 39) = 1.575$, $p\text{-value} = 0.079$, reflecting the substantial heterogeneity typical of cognitive outcomes in dementia populations. Sequence could not be formally evaluated in this ANOVA table because its denominator mean square is provided by the subject-within-sequence term; however, earlier diagnostic tests already showed no evidence of a sequence effect.

Overall, the classical crossover ANOVA shows that treatment-related variability is small relative to the natural variability across individuals. After accounting for period and sequence effects, there is no statistical evidence that real tACS produced greater short-term improvement in ADAS-Cog performance than sham stimulation.

2.5.4 ANCOVA Adjusting for Baseline Differences

In crossover studies, baseline adjustment is critical because individuals often differ substantially in their cognitive performance before receiving any treatment. Such between-subject variability inflates residual noise and can obscure true treatment differences. To address this, an ANCOVA was applied, with post-treatment ADAS-Cog score as the dependent variable, treatment (real vs. sham tACS) as the primary factor, the corresponding baseline ADAS-Cog score as a covariate, and period (1 vs. 2) included to adjust for temporal effects. This approach follows the widely recommended practice in clinical trials, where ANCOVA is preferred to simple change-score analyses because it provides greater statistical power and yields less biased estimates of treatment effects [83]. The fitted model can be expressed as:

$$\text{Treated}_{ij} = \beta_0 + \beta_1 \text{Treatment}_{ij} + \beta_2 \text{Baseline}_{ij} + \beta_3 \text{Period}_j + \varepsilon_{ij}.$$

In this model, β_1 represents the adjusted treatment effect and tests whether real tACS differs from sham after accounting for baseline and period. The coefficient β_2 captures the influence of the period-specific baseline score on the post-treatment outcome, while β_3 represents any systematic difference between Period 1 and Period 2. The intercept β_0 is not of inferential interest. Thus, the hypothesis of primary interest is

$$H_0 : \beta_1 = 0,$$

which evaluates whether real stimulation produces a different post-treatment ADAS-Cog score than sham stimulation once baseline variability is controlled for.

The ANCOVA model demonstrated excellent explanatory power, accounting for 80.7% of the variance in post-treatment ADAS-Cog scores ($R^2 = 0.807$, adjusted $R^2 = 0.800$), and was highly significant overall, $F(3, 78) = 108.9$, $p\text{-value} < 0.001$. Baseline ADAS-Cog was by far the strongest predictor of the post-treatment score ($\beta = 0.93$, $SE = 0.05$, $t = 18.02$, $p\text{-value} < 0.001$), confirming that cognitive impairment levels are highly stable within individuals and vary largely between subjects. This strong baseline dependence illustrates why ANCOVA is advantageous: by adjusting for baseline, it removes much of the between-subject noise, reduces residual variance, and provides more precise estimates of the treatment effect.

In contrast, neither period nor treatment contributed significantly to the model. The period effect was negligible ($\beta = 0.06$, $SE = 1.01$, $p\text{-value} = 0.95$), indicating no systematic difference between Period 1 and Period 2 after adjusting for baseline and treatment. The adjusted treatment effect comparing real with sham tACS was also not statistically significant ($\beta = -0.36$, $SE = 1.01$, $t = -0.35$, $p\text{-value} = 0.72$). Although the coefficient direction suggested slightly lower (improved) ADAS-Cog scores following real stimulation, the corresponding confidence interval (-2.37 to 1.65) spanned both meaningful improvement and worsening, indicating that any true treatment effect is likely to be small.

Taken together, the ANCOVA results suggest that, while baseline performance strongly

predicts post-treatment outcomes, real tACS did not produce a statistically detectable improvement relative to sham stimulation in this sample.

2.6 Conclusion

The classical statistical analyses examined treatment, period, and sequence effects within the 2×2 crossover design, beginning with checks of key design assumptions. No evidence of sequence imbalance, baseline differences, or unequal variances was found, supporting the validity of using both periods in the analysis. Paired t-tests and between-sequence comparisons showed small improvements following both real and sham stimulation, but none of these differences were statistically significant. The classical crossover ANOVA confirmed these findings, indicating that treatment-related variability was small relative to the substantial natural variability among participants.

Adjusting for baseline cognitive status through ANCOVA further clarified the structure of the data. Baseline ADAS-Cog scores were by far the strongest predictor of post-treatment performance, while neither treatment nor period showed significant effects after adjustment. Together, these classical methods indicate that real tACS did not produce measurable short-term improvement in ADAS-Cog scores compared with sham stimulation. These results highlight the considerable individual variability in cognitive outcomes and motivate the use of additional methods, such as bootstrap analysis, to better characterize uncertainty in treatment effects.

Chapter 3

Bootstrap analyses of 2×2 crossover studies

Summary. This chapter introduces the bootstrap as a flexible alternative to classical parametric and nonparametric methods for inference in medical research. We first outline how resampling from the empirical distribution can approximate the sampling distribution of an estimator, enabling the construction of confidence intervals and hypothesis tests without strong distributional assumptions. We then describe variants of the bootstrap relevant to data arising from our two-period, two-treatment crossover study, developing bootstrap-based procedures for treatment comparison, crossover ANOVA, and ANCOVA with baseline adjustment.

3.1 Introduction

In medical research, we often aim to draw conclusions about a broader population from a single sample. Estimates such as means, differences in proportions, or treatment effects are commonly accompanied by confidence intervals based on standard errors, intended to quantify the uncertainty that would arise under repeated sampling from the population of interest

[84]. Because we usually observe only one sample, however, the accuracy of these intervals depends critically on assumptions about the underlying data-generating distribution. Parametric methods address this problem by assuming a known form for the distribution (most commonly normality), which facilitates analytical derivation of standard errors and test statistics, especially in large samples. In practice, these assumptions may be questionable (as we observed in Chapter 2). Medical and psychological data are frequently skewed, bounded, ordinal, or based on small samples, leading to violations of normality and homoscedasticity. Under such conditions, parametric procedures can yield misleading or invalid inferences [85]. Nonparametric methods, such as the Wilcoxon signed-rank test or the Mann–Whitney U test, relax distributional assumptions and are therefore attractive when dealing with non-normal or ordinal data. However, they are typically geared toward hypothesis testing and primarily provide p -values, rather than direct estimates of effect sizes and associated confidence intervals [86]. As a result, while non-parametric tests are robust, they may be less informative in settings where quantifying the magnitude and uncertainty of treatment effects is essential.

To bridge this gap between parametric precision and non-parametric robustness, Bradley Efron introduced the bootstrap in 1979 [87]. The central idea is to approximate the sampling distribution of an estimator by resampling from the observed data. If the sample adequately represents the population, then resampling with replacement from it can mimic repeated sampling from the underlying distribution. This approach allows estimation of standard errors, construction of confidence intervals, and hypothesis testing without strong parametric assumptions. Efron and Tibshirani later formalized and extended this framework, demonstrating that the bootstrap can be applied to a wide range of statistical problems with minimal modeling assumptions [88]. Refinements such as the bias-corrected and accelerated (BCa) interval further improve accuracy in the presence of skewed or non-symmetric sampling distributions [89].

Since its introduction, the bootstrap has become widely used in areas where data

are noisy, irregular, or limited in size. In medical research, applications include assessing the stability of odds ratios in logistic regression, comparing survival curves in oncology, and evaluating changes in scores from cognitive instruments [90, 91]. These methods are particularly relevant in neurology and psychiatry, where outcomes such as the Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS–Cog) are bounded and often non-normal. The flexibility of the bootstrap is especially valuable in small clinical trials, where traditional parametric assumptions are tenuous and robust estimation is needed [92].

In this chapter we apply the bootstrap method to our crossover study. To this end, we first provide a high level introduction to the Bootstrap approach and its application in approximating the sampling distribution, constructing confidence interval and performing testing hypothesis. We then provide some variants of the bootstrap technique that are relevant to our study. Finally we provide bootstrap tests for treatment comparison, crossover ANOVA and ANCOVA tests.

3.2 A high level introduction to bootstrap

At a high level, the bootstrap is a data-driven method for approximating what would happen if a study could be repeated many times, while only observing one dataset. Classical approaches answer this question by assuming a parametric model (e.g., normal distribution) and deriving the sampling distribution of a statistic from that model. The bootstrap instead treats the observed sample as a proxy for the unknown population and approximates repeated sampling by resampling from the data themselves. This idea underlies both estimation (standard errors and confidence intervals) and hypothesis testing, where in each case we approximate the distribution of a statistic by recomputing it many times on resampled versions of the data.

3.2.1 Bootstrap as an approximation to the sampling distribution

Consider a sample $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$ drawn from an unknown population distribution F . Let $\theta = t(F)$ denote the parameter of interest, such as a mean, median, or regression coefficient, and let $\hat{\theta}$ denote the estimator computed from the observed data. The sampling distribution of $\hat{\theta} = g(\mathbf{X}) = g(\mathbf{X}_1, \dots, \mathbf{X}_n)$ under F is typically unknown and analytically intractable. The nonparametric bootstrap replaces F with the empirical distribution $\hat{F}_n$, which assigns mass $1/n$ to each observed data point. In the bootstrap world,

$$\mathbf{X}^* = \{\mathbf{X}_1^*, \dots, \mathbf{X}_n^*\} \stackrel{\text{iid}}{\sim} \hat{F}_n,$$

and the bootstrap version of the estimator is $\hat{\theta}^* = g(\mathbf{X}_1^*, \dots, \mathbf{X}_n^*)$. Repeating this resampling B times yields bootstrap replicates $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ drawn from the bootstrap distribution, i.e., the conditional distribution of $\hat{\theta}^*$ given the data. The classical nonparametric bootstrap proceeds as follows:

1. Generate B bootstrap samples $\mathbf{X}^{*1}, \mathbf{X}^{*2}, \dots, \mathbf{X}^{*B}$, each of size n , by sampling with replacement from $\mathbf{X}$.
2. For each bootstrap sample $\mathbf{X}^{*b}$, compute the statistic $\hat{\theta}^{*b} = g(\mathbf{X}^{*b})$.
3. Use the empirical distribution of $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B}$ to approximate the sampling distribution of $\hat{\theta}$.

Under mild regularity conditions, the bootstrap distribution converges (in an appropriate sense) to the true sampling distribution, so that

$$\sup_t \left| \mathbb{P}^*(\hat{\theta}^* \leq t \mid \mathbf{X}) - \mathbb{P}(\hat{\theta} \leq t) \right| \xrightarrow{P} 0.$$

This justifies the bootstrap as a consistent estimator of the sampling distribution of $\hat{\theta}$. Also, the standard error of $\hat{\theta}$ is estimated as

$$\text{SE}_{\text{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B \left(\hat{\theta}^{*b} - \bar{\hat{\theta}}^* \right)^2},$$

where $\bar{\hat{\theta}}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{*b}$ denotes the mean of the bootstrap replicates. The empirical distribution of $\hat{\theta}^{*b}$ can then be used to construct confidence intervals (e.g., percentile or BCa), estimate bias, and perform hypothesis testing without strong distributional assumptions.

3.2.2 Bootstrap hypothesis testing

In classical testing, we consider a null hypothesis H_0 about F (for example, equality of means in two groups) and construct a test statistic $T_n = T(\mathbf{X}_1, \dots, \mathbf{X}_n)$ whose distribution under H_0 is assumed known or approximated asymptotically. A p -value is then computed as

$$p\text{-value} = \mathbb{P}_{H_0}(|T_n| \geq |T_{\text{obs}}|),$$

where T_{obs} is the observed value of the statistic. When the exact null distribution is unknown or the sample size is small, such approximations may be unreliable. The bootstrap provides a resampling-based alternative by approximating the null distribution of T_n ,

$$G_{0,n}(t) = \mathbb{P}_{H_0}(T_n \leq t),$$

with a bootstrap distribution $G_{0,n}^*(t \mid \mathbf{X})$ constructed from resampled data that satisfy H_0 .

A general recipe is:

1. **Choose a test statistic.** Specify a statistic $T_n = T(\mathbf{X}_1, \dots, \mathbf{X}_n)$ that measures evidence against H_0 (e.g., difference in means, a regression coefficient, an F -statistic)

and compute T_{obs} .

2. **Construct a “null world.”** Modify or recenter the data so that H_0 holds by construction, for example by:

- centering outcomes so that group means are equal under H_0 ,
- fitting a reduced model without the effect being tested and using fitted values plus resampled residuals,
- in two-sample problems, pooling data and resampling without labels when H_0 states that the two groups share a common distribution.

3. **Generate bootstrap samples under H_0 .** From this null world, generate B bootstrap datasets $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ using an appropriate structure-preserving scheme (e.g., subject-level or residual resampling).

4. **Recompute the test statistic.** For each bootstrap dataset, compute $T_n^{*b} = T(\mathbf{X}^{*b})$ for $b = 1, \dots, B$. The empirical distribution of $T_n^{*1}, \dots, T_n^{*B}$ approximates the null distribution of T_n .

5. **Compute a bootstrap p -value.** For a two-sided test, p -values is given by $p_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(|T_n^{*b}| \geq |T_{\text{obs}}|)$, i.e. the proportion of bootstrap test statistics that are at least as extreme as the observed statistic. One-sided tests are defined analogously.

Thus, bootstrap hypothesis testing can be viewed as a plug-in approximation to the ideal, but unknown, null distribution of the test statistic. In this approach, the true F is replaced by an empirical or model-based estimate that enforces H_0 , and the resulting conditional distribution of T_n^* is used to quantify evidence against the null.

3.2.3 Variants of the bootstrap and relevance for this study

While this study focuses on the nonparametric bootstrap due to its minimal assumptions and broad applicability, several variants have been developed for specific data structures:

- Residual bootstrap: resampling residuals in regression settings [93].
- Wild bootstrap: designed for heteroscedastic errors via randomly weighted residuals [94].
- Block bootstrap: used for dependent data such as time series or spatial observations [95].
- BCa bootstrap: a refinement that adjusts for bias and skewness to construct more accurate intervals [89].

These adaptations address particular challenges, but the classical nonparametric bootstrap remains especially suitable for the present crossover study, where cognitive questionnaire data are bounded and may deviate substantially from normality. In addition to estimation, the general testing framework above is particularly valuable when parametric assumptions are doubtful or sample sizes are modest. Traditional hypothesis tests rely on theoretical null distributions such as the t -distribution, which may be inappropriate for bounded, ordinal, or skewed data common in biomedical research. By approximating the null distribution via resampling under H_0 and computing bootstrap-based p -values, the bootstrap accommodates complex designs and non-normal outcomes.

In the context of our small-sample crossover trial using the ADAS–Cog instrument, the nonparametric bootstrap therefore provides a robust alternative to standard parametric methods, enabling valid inference on treatment effects without imposing strict distributional assumptions [88, 89].

3.3 Bootstrap Treatment Comparison

In the classical analysis, the between-sequence comparison of treatment-response was based on the two-sample t -statistic applied to the subject-level difference scores

$$D = \Delta_{\text{Real}} - \Delta_{\text{Sham}},$$

with D_{AB} and D_{BA} defined as in the previous section. In the present dataset, there were $n_{\text{AB}} = 19$ subjects in the AB sequence and $n_{\text{BA}} = 22$ subjects in the BA sequence. The corresponding sample means of the difference scores were $\bar{D}_{\text{AB}} = -0.5747$ and $\bar{D}_{\text{BA}} = -0.2305$.

Let $\bar{D}_{\text{AB}}$ and $\bar{D}_{\text{BA}}$ denote the sample means, and let

$$\hat{\delta} = \bar{D}_{\text{AB}} - \bar{D}_{\text{BA}}$$

be the observed difference in mean treatment response between the two sequences. For the observed data, this quantity was

$$\hat{\delta} = -0.3443.$$

The usual pooled-variance t -statistic can be written as

$$T = \frac{\hat{\delta}}{\hat{\sigma}_p \sqrt{\frac{1}{n_{\text{AB}}} + \frac{1}{n_{\text{BA}}}}},$$

where $\hat{\sigma}_p^2$ is the pooled within-sequence variance and n_{AB} , n_{BA} are the sequence sample sizes.

Using the pooled standard deviation estimate $\hat{\sigma}_p = 5.7074$, the observed test statistic was

$$T = -0.1926.$$

3.3.1 Nonparametric subject-level bootstrap

Because the unit of randomization is the subject, we construct a nonparametric subject-level bootstrap that resamples difference scores within each sequence. Let $D_{AB,1}, \dots, D_{AB,n_{AB}}$ and $D_{BA,1}, \dots, D_{BA,n_{BA}}$ denote the observed difference scores.

For $b = 1, \dots, B$ bootstrap replicates:

1. Resample with replacement n_{AB} subjects from $\{D_{AB,1}, \dots, D_{AB,n_{AB}}\}$. This will result in $D_{AB,1}^{*(b)}, \dots, D_{AB,n_{AB}}^{*(b)}$.
2. Similarly, resample with replacement n_{BA} subjects from $\{D_{BA,1}, \dots, D_{BA,n_{BA}}\}$ to obtain $D_{BA,1}^{*(b)}, \dots, D_{BA,n_{BA}}^{*(b)}$.
3. Compute the bootstrap mean difference $\hat{\delta}^{*(b)} = \bar{D}_{AB}^{*(b)} - \bar{D}_{BA}^{*(b)}$, and the corresponding pooled-variance t -statistic $T^{*(b)}$ in the same way as T .

The empirical distribution of $\hat{\delta}^{*(b)}$ approximates the sampling distribution of $\hat{\delta}$, and the empirical distribution of $T^{*(b)}$ approximates the sampling distribution of T under the fitted model.

3.3.2 Bootstrap standard error, confidence interval, and test

The bootstrap standard error of $\hat{\delta}$ is

$$\widehat{\text{se}}_{\text{boot}}(\hat{\delta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\delta}^{*(b)} - \bar{\delta}^*)^2},$$

with $\bar{\delta}^* = \frac{1}{B} \sum_{b=1}^B \hat{\delta}^{*(b)}$. For the present dataset, the mean of the 1000 bootstrap replicates ($B = 1000$) was $\bar{\delta}^* = -0.3630$, and the resulting bootstrap standard error was $\widehat{\text{se}}_{\text{boot}}(\hat{\delta}) = 1.7142$.

A basic $(1 - \alpha)$ percentile bootstrap confidence interval for δ is obtained by taking the empirical $\alpha/2$ and $1 - \alpha/2$ quantiles of the set $\{\hat{\delta}^{*(1)}, \dots, \hat{\delta}^{*(B)}\}$:

$$[\hat{\delta}_{(\alpha/2)}^*, \hat{\delta}_{(1-\alpha/2)}^*].$$

Using $\alpha = 0.05$, the 95% percentile bootstrap interval was

$$[-3.7151, 2.9841].$$

More accurate intervals can be constructed using bias-corrected and accelerated (BCa) methods if desired.

To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$, a bootstrap test that respects the null can be defined by first centering the difference scores within each sequence,

$$\tilde{D}_{AB,j} = D_{AB,j} - \bar{D}_{AB}, \quad \tilde{D}_{BA,j} = D_{BA,j} - \bar{D}_{BA},$$

and then resampling from the centered pooled set $\{\tilde{D}_{AB,j}, \tilde{D}_{BA,j}\}$ under the constraint of equal means. Repeating the above bootstrap procedure with centered scores yields a null bootstrap distribution $\{T_0^{*(b)}\}$, from which a two-sided bootstrap p -value can be computed as

$$p_{\text{boot}} = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(|T_0^{*(b)}| \geq |T_{\text{obs}}|).$$

For the present data, this test yielded a bootstrap p -value of $p_{\text{boot}} = 0.843$. The bootstrap analysis indicates that the observed mean difference $\hat{\delta}$ is entirely consistent with sampling variability. The large bootstrap standard error, the wide confidence interval spanning zero, and the very large bootstrap p -value ($p = 0.843$) all indicate that there is no evidence of a difference in mean treatment response between the two sequences based on the subject-

level difference scores. Under regularity conditions, the subject-level bootstrap consistently approximates the sampling distribution of $\hat{\delta}$ and of the test statistic T , thereby providing an alternative inference procedure that does not rely on normality assumptions for the difference scores.

3.3.3 Bootstrap Crossover ANOVA

The classical 2×2 crossover ANOVA in Chapter 2 is based on the linear model

$$\Delta_{ijk} = \mu + \text{Sequence}_i + \text{Subject}(\text{Sequence})_{ij} + \text{Period}_k + \text{Treatment}_{ik} + \varepsilon_{ijk},$$

with i indexing sequence (AB or BA), j indexing subjects within sequence, and k indexing period. The treatment, period, and other fixed effects are tested using F -statistics that compare appropriate mean squares. In small samples, the exact distribution of these F -statistics may deviate from the nominal F distribution, especially when normality or homoscedasticity assumptions are questionable. A nonparametric cluster bootstrap provides an alternative way to approximate the sampling distribution of the treatment and period test statistics while preserving the crossover structure.

3.3.4 Cluster (subject-level) bootstrap scheme

For each sequence i , the basic observational unit is the subject, who contributes a vector of responses $\Delta_{ij} = (\Delta_{ij1}, \Delta_{ij2})^\top$, together with the corresponding design indicators (sequence, period, treatment). A cluster bootstrap resamples these subject-level vectors within each sequence. For $b = 1, \dots, B$:

1. For each sequence, resample with replacement n_i subjects from $\{\Delta_{i1}, \dots, \Delta_{in_i}\}$ to obtain $\Delta_{i1}^{*(b)}, \dots, \Delta_{in_i}^{*(b)}$, where $i \in \{\text{AB}, \text{BA}\}$.

2. For the resulting bootstrap sample, fit the same crossover ANOVA model as in the classical analysis, using the original design structure (sequence, period, treatment, and subject-within-sequence).
3. Record the bootstrap F -statistics $F_{\text{trt}}^{*(b)}$, $F_{\text{per}}^{*(b)}$, etc. for the effects of interest (e.g., treatment and period).

The empirical distributions of $\{F_{\text{trt}}^{*(b)}\}$ and $\{F_{\text{per}}^{*(b)}\}$ approximate the sampling distributions of the corresponding F -statistics, taking into account the random variation in the subject-specific trajectories without imposing a parametric form on the error distribution.

3.3.5 Bootstrap inference for treatment and period effects

To evaluate the significance of the treatment and period effects in the crossover ANOVA model, the observed F -statistics were compared with their bootstrap distributions. Let $F_{\text{tret,obs}}$ and $F_{\text{per,obs}}$ denote the observed F -statistics from the original data. For the present dataset, the classical ANOVA yielded

$$F_{\text{tret,obs}} = 0.203, \quad F_{\text{per,obs}} = 0.037,$$

both of which are small, indicating little evidence for either effect under the classical model.

To obtain inference without relying on normality or homoscedasticity assumptions, a subject-level cluster bootstrap was implemented. Entire subject profiles were resampled within sequence, preserving the AB/BA allocation and the within-subject correlation. For each bootstrap sample, the same crossover ANOVA model was fitted and the bootstrap F -statistics $\{F_{\text{tret}}^{*(b)}\}$ and $\{F_{\text{per}}^{*(b)}\}$ were recorded. With $B = 1000$ valid bootstrap replications, these sets provided empirical approximations to the sampling distributions of the treatment and period test statistics.

Two-sided bootstrap p -values were computed as

$$p_{\text{tret},\text{boot}} = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(F_{\text{tret}}^{*(b)} \geq F_{\text{tret},\text{obs}}), \quad p_{\text{per},\text{boot}} = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(F_{\text{per}}^{*(b)} \geq F_{\text{per},\text{obs}}).$$

For this analysis, the bootstrap tests yielded

$$p_{\text{tret},\text{boot}} = 0.697, \quad p_{\text{per},\text{boot}} = 0.844.$$

Both values are large, indicating that the observed statistics fall well within the range expected under the null hypothesis. Thus, the bootstrap results reinforce the classical ANOVA conclusion: within this small-sample 2×2 crossover trial, neither the treatment effect nor the period effect shows evidence of being different from zero. The bootstrap analysis therefore provides distribution-free support for the absence of a detectable difference in ADAS–Cog change between real and sham tACS.

Although a residual-based bootstrap can be used to obtain a null-constrained distribution of F_{tret} , the present analysis focuses on the nonparametric subject-level bootstrap, which directly resamples subject clusters and preserves the crossover structure. The reported results therefore correspond to the nonparametric bootstrap procedure described above.

3.3.6 Bootstrap ANCOVA with Baseline Adjustment

The ANCOVA model in Chapter 2 relates post-treatment ADAS–Cog scores Treated_{ij} to treatment, baseline, and period:

$$\text{Treated}_{ij} = \beta_0 + \beta_1 \text{Treatment}_{ij} + \beta_2 \text{Baseline}_{ij} + \beta_3 \text{Period}_j + \varepsilon_{ij},$$

where β_1 captures the adjusted effect of real versus sham tACS. Repeated measurements within subjects induce correlation in the errors ε_{ij} , and the small sample size may limit the

reliability of standard normal-theory inference for $\hat{\beta}_1$. A bootstrap approach can again be constructed at the subject level, either by resampling subjects nonparametrically or by using a residual cluster bootstrap.

In the present dataset, the classical ANCOVA yielded an adjusted treatment effect of $\hat{\beta}_1 = -0.358$ with standard error $SE(\hat{\beta}_1) = 1.010$. The corresponding t -statistic was $T_{\beta_1, \text{obs}} = -0.354$, and the 95% confidence interval for β_1 was $[-2.369, 1.653]$. These results indicate no detectable difference between the real and sham tACS conditions after adjusting for baseline and period effects, motivating the use of bootstrap procedures to obtain inference that does not rely on normal-theory assumptions.

3.3.7 Nonparametric subject-level bootstrap

Each subject i contributes a small vector of observations

$$(\text{Treated}_{i1}, \text{Baseline}_{i1}, \text{Treatment}_{i1}, \text{Period}_1; \text{Treated}_{i2}, \text{Baseline}_{i2}, \text{Treatment}_{i2}, \text{Period}_2).$$

For $b = 1, \dots, B$:

1. Resample with replacement N subjects from the set of N original subjects (across both sequences).
2. For the resulting bootstrap sample, fit the same ANCOVA model as above by ordinary least squares.
3. Record the bootstrap estimates $\hat{\beta}_1^{*(b)}$, and if desired, their associated t -statistics $T_{\beta_1}^{*(b)}$.

The empirical distribution of $\{\hat{\beta}_1^{*(b)}\}$ approximates the sampling distribution of $\hat{\beta}_1$, with the within-subject correlation preserved by resampling entire subject profiles.

The nonparametric subject-level bootstrap was applied using $B = 1000$ replicates, with resampling performed at the level of the subject to preserve the within-subject depen-

dence inherent in the crossover design. The empirical bootstrap distribution of $\{\hat{\beta}_1^{*(b)}\}$ was centered near the classical ANCOVA estimate, with a bootstrap mean of $\bar{\beta}_1^* = -0.328$. The bootstrap standard error was $\widehat{\text{se}}_{\text{boot}}(\hat{\beta}_1) = 0.846$, slightly smaller than the normal-theory estimate from the classical model. A percentile 95% bootstrap confidence interval for the adjusted treatment effect was $[-1.995, 1.287]$, which again includes zero and indicates no detectable treatment effect after adjusting for baseline and period. These results demonstrate that the bootstrap-based sampling distribution of $\hat{\beta}_1$ closely agrees with the classical estimate while avoiding reliance on normality assumptions.

3.3.8 Residual cluster bootstrap under the null

For testing $H_0 : \beta_1 = 0$, it is often preferable to construct a bootstrap distribution that respects the null hypothesis. This can be done via a residual cluster bootstrap:

1. Fit the reduced ANCOVA model without the treatment term (i.e., set $\beta_1 = 0$) to obtain fitted values $\widehat{\text{Treated}}_{ij}^{(0)}$ and residuals $\hat{\varepsilon}_{ij}^{(0)}$ for each subject-period pair.
2. For each subject i , collect the residual vector $\hat{\boldsymbol{\varepsilon}}_i^{(0)} = (\hat{\varepsilon}_{i1}^{(0)}, \hat{\varepsilon}_{i2}^{(0)})^\top$.
3. For $b = 1, \dots, B$, resample with replacement N residual vectors $\hat{\boldsymbol{\varepsilon}}_{i^{*(b)}}^{(0)}$ across subjects, and construct bootstrap responses under H_0 via

$$\text{Treated}_{ij}^{*(b)} = \widehat{\text{Treated}}_{ij}^{(0)} + \hat{\varepsilon}_{i^{*(b)}j}^{(0)}.$$

The covariates Treatment_{ij} , Baseline_{ij} , and Period_j remain fixed at their observed values.

4. Fit the full ANCOVA model to each bootstrap dataset to obtain $\hat{\beta}_1^{*(b)}$ and the corresponding t -statistic $T_{\beta_1}^{*(b)}$.

The resulting $\{T_{\beta_1}^{*(b)}\}$ approximate the null distribution of the treatment t -statistic.

A two-sided bootstrap p -value is then

$$p_{\beta_1, \text{boot}} = \frac{1}{B} \sum_{b=1}^B \mathbb{I}(|T_{\beta_1}^{*(b)}| \geq |T_{\beta_1, \text{obs}}|),$$

where $T_{\beta_1, \text{obs}}$ is the observed t -statistic from the original data. Percentile or BCa intervals for β_1 can be constructed from the empirical distribution of $\{\hat{\beta}_1^{*(b)}\}$. Under standard regularity conditions (smoothness of the regression functional, bounded moments, and stable cluster sizes), the subject-level and residual cluster bootstraps consistently approximate the sampling distribution of $\hat{\beta}_1$ and its associated test statistic, providing a robust alternative to normal-theory ANCOVA inference in small-sample crossover settings.

To obtain a null distribution that respects the hypothesis $H_0 : \beta_1 = 0$, a residual cluster bootstrap was performed following the procedure described above. The reduced ANCOVA model without the treatment term was first fitted to the data, and subject-level residual vectors $(\hat{\varepsilon}_{i1}^{(0)}, \hat{\varepsilon}_{i2}^{(0)})$ were extracted together with the fitted values from the reduced model. For each bootstrap replication, a residual vector was sampled with replacement across subjects and added to the reduced-model fitted values, while all covariates (baseline, treatment indicator, and period) were held fixed. This produced bootstrap responses that conformed to H_0 . The full ANCOVA model was then refitted to each bootstrap dataset, yielding a bootstrap t -statistic $T_{\beta_1}^{*(b)}$ for the treatment effect.

The resulting null distribution of $\{T_{\beta_1}^{*(b)}\}$ was centered near zero, as expected under H_0 , with representative values such as $(-0.587, -0.013, -0.029, 0.439, 0.018)$. The observed t -statistic from the original ANCOVA was $T_{\beta_1, \text{obs}} = -0.354$. Comparing the observed value to the bootstrap null distribution gave a two-sided bootstrap p -value of $p_{\beta_1, \text{boot}} = 0.65$, indicating that statistics as or more extreme than $T_{\beta_1, \text{obs}}$ occur frequently under the null model. Thus, the residual cluster bootstrap provides no evidence against $H_0 : \beta_1 = 0$, and confirms that after adjusting for baseline and period, the treatment effect is not statistically distinguishable from zero.

3.4 Conclusion

The bootstrap analyses presented in this chapter provided an assumption-light assessment of treatment, period, and adjusted treatment effects in the crossover trial. Across all resampling approaches, the estimated treatment effects remained small and unstable. The subject-level bootstrap for the treatment–response difference produced a 95% percentile interval of $[-3.72, 2.98]$, which spans zero and indicates no detectable difference between real and sham stimulation. Similarly, the ANCOVA bootstrap yielded a bootstrap mean adjusted treatment effect of $\bar{\beta}_1^* = -0.328$ with a bootstrap standard error of 0.846 and a percentile 95% confidence interval of $[-1.995, 1.287]$, again demonstrating that the distribution of plausible treatment effects includes zero and provides no evidence of benefit.

The null-based residual cluster bootstrap further reinforced this conclusion. The observed ANCOVA t -statistic for the treatment term ($T_{\beta_1, \text{obs}} = -0.354$) fell comfortably within the null bootstrap distribution, yielding a large bootstrap p -value of 0.65. This indicates that statistics of similar or greater magnitude arise frequently under the assumption of no treatment effect. Taken together, the classical analyses and all bootstrap variants, including subject-level, cluster, and residual-based bootstraps, consistently support the same finding: real tACS did not produce a measurable improvement in ADAS–Cog performance relative to sham stimulation in this small-sample crossover dataset. The bootstrap framework therefore confirms the robustness of the classical results while providing an inference approach that does not rely on strong distributional assumptions.

Chapter 4

Bayesian analyses of 2×2 crossover studies

Summary. This chapter presents a Bayesian framework for a 2×2 crossover trials with baseline measurements before each period. Using a compound-symmetric covariance to capture within-subject correlation, we derive closed-form posteriors for means and variance components, identify simple sufficient statistics with chi-square structure, and provide Bayes factors to handle uncertainty about period and carryover effects. The approach delivers direct, clinically interpretable probabilities for treatment contrasts, with practical guidance on prior choice, model averaging, and computation.

4.1 Introduction

Bayesian analysis is particularly attractive for two-period, two-treatment crossover designs with baseline measurements before each treatment period. Compared to traditional frequentist analyses, Bayesian models (i) accommodate hierarchical structure and correlation within subjects, (ii) produce direct probability statements for clinically meaningful contrasts (e.g., $\mathbb{P}(\tau > \Delta \mid Y)$), and (iii) permit principled model comparison and model averaging when pe-

riod or carryover effects are uncertain; see Senn [2], Grieve [96], Gelman et al. [97]. This chapter develops a coherent Bayesian formulation for a two-sequence design (AB and BA) with two baselines and two post-treatment outcomes per subject, using a compound-symmetric covariance for within-subject measurements. We give clean posteriors for means and variance components, identify convenient sufficient statistics (SS_1, SS_2) with chi-square structure, and outline Bayes factors using proper priors (with a pragmatic alternative noted).

Traditional analyses in clinical research have been dominated by the frequentist framework, where probability is interpreted as long-run frequency under repeated sampling. In that setting, hypothesis testing plays a central role and the p -value is often used as a primary measure of evidence. Yet the p -value is frequently misinterpreted: it is not the probability that the null hypothesis is true and it depends strongly on sample size, so trivial effects may appear significant in large samples while meaningful effects may fail to reach significance in small samples [98]. Moreover, p -values refer to unobserved, more-extreme data rather than to the evidence contained in the data actually observed, a feature at odds with the likelihood principle [99, 100].

Bayesian methods address these limitations by treating unknown parameters as random variables and updating prior beliefs with the observed data through Bayes theorem [97, 101]. The resulting posterior distribution provides direct probability statements about parameters of interest (e.g., the probability the treatment effect exceeds a clinically relevant threshold), which aligns with how evidence is interpreted in practice [102, 103]. When data are abundant, the likelihood dominates and Bayesian and frequentist conclusions often agree; when samples are small, a common situation in crossover studies, incorporating prior information can improve stability and efficiency.

In crossover trials, within-subject comparisons reduce between-subject variability, but the analysis must still contend with correlation among repeated measures, possible period effects, and potential carryover from the first to the second period. Bayesian modeling

accommodates these features within a single framework and allows uncertainty about period and carryover effects to be handled explicitly via model comparison and model averaging. In particular, Bayes factors quantify the relative support for competing models without relying on arbitrary significance thresholds, and model averaging propagates structural uncertainty into the final inference [103, 99].

The design considered here is the standard two-sequence, two-period, two-treatment crossover with sequences AB and BA. Each subject contributes two baseline measurements (one before each period) and two post-treatment outcomes. We adopt a compound-symmetric working covariance to capture the dominant intraclass correlation while retaining analytic tractability. Building on Grieve [96] and Senn [2], we show that sufficient statistics (SS_1, SS_2) with scaled chi-square structure lead to conjugate posteriors for mean parameters and variance components. This facilitates posterior summaries for treatment, period, and carryover effects, as well as Bayes factors under proper priors. Where closed-form expressions are inconvenient, a pragmatic alternative is noted.

From a computational perspective, while early barriers to Bayesian adoption stemmed from analytical intractability, standard Markov chain Monte Carlo (MCMC) methods now make posterior computation routine when closed forms are unavailable [104]. Thus the proposed analysis can proceed either fully analytically (under CS with conjugate priors) or with straightforward simulation-based updates, while preserving the interpretability of posterior probabilities for clinically meaningful contrasts.

The rest of this chapter is organized as follow. We first specify the joint likelihood for the two-sequence, two-period design with two baselines and two post-treatment outcomes per subject under a CS working covariance. We then identify the sufficient statistics (SS_1, SS_2) and derive conjugate posteriors for means and variance components. Next, we discuss prior choices (including proper priors for Bayes factors) and develop Bayes factors for treatment, period, and carryover hypotheses, with a pragmatic alternative. We conclude with guidance

on reporting posterior probabilities for clinically relevant contrasts and with brief comments on computation and implementation [105, 106, 107, 108].

4.2 Bayesian Framework for Crossover Study

Baseline measurements are collected during non-treatment periods and provide information that enhances the estimation of direct treatment effects [2]. These measurements are typically taken before the start of each treatment phase, and much of the methodological literature on their use pertains to two-period crossover designs. In the present study, baseline data were gathered prior to each treatment period and are incorporated into a Bayesian analysis to improve the precision and robustness of the estimated treatment effects.

4.2.1 Models in the two-period crossover with baselines

Several models have been proposed to incorporate baseline measurements into two-period crossover designs (e.g., Willan and Pater [109], Varma and Chilton [110], Kenward and Jones [111], Chi [112], and Patel [113]). Although these models differ in structure, they can generally be expressed in terms of the cell means corresponding to each treatment, period, and sequence combination. Let y_{ijk} denote the outcome for the j th subject in sequence $i \in \{AB, BA\}$ during period k . The associated expected value (cell mean) is $\mu_{ik} = E(y_{ijk})$. Crossover models are then specified by additive effects applied to the cell means, typically including treatment, period, sequence, carryover, and (implicitly) the role of baselines. In a two-period crossover with baselines collected at the start of each period, each participant contributes four observations corresponding to four time points: the initial run-in (pre-treatment) period, the first treatment phase, the washout interval, and the second treatment phase.

Table 4.1: Cell mean structures under various crossover models.

Sequence	Measurement	Period 1	Period 2	Model
A → B	Pre-treatment	$\mu + \pi$	$\mu - \pi + \lambda$	(a) Willan and Pater's Model
	Post-treatment	$\mu + \pi + \tau$	$\mu - \pi - \tau + \lambda$	
B → A	Pre-treatment	$\mu + \pi$	$\mu - \pi - \lambda$	
	Post-treatment	$\mu + \pi - \tau$	$\mu - \pi + \tau - \lambda$	
A → B	Pre-treatment	$\mu + \pi$	$\mu - \pi + \theta$	(b) Varma and Chilton's model
	Post-treatment	$\mu + \pi + \tau$	$\mu - \pi - \tau + \lambda$	
B → A	Pre-treatment	$\mu + \pi$	$\mu - \pi - \theta$	
	Post-treatment	$\mu + \pi - \tau$	$\mu - \pi + \tau - \lambda$	
A → B	Pre-treatment	$\mu + \gamma + \pi_1$	$\mu + \gamma + \pi_3 + \lambda$	(c) Chi's model
	Post-treatment	$\mu + \gamma + \pi_2 + \tau$	$\mu + \gamma - \pi_1 - \pi_2 - \pi_3 - \tau + \lambda$	
B → A	Pre-treatment	$\mu - \gamma + \pi_1$	$\mu - \gamma + \pi_3 - \lambda$	
	Post-treatment	$\mu - \gamma + \pi_2 - \tau$	$\mu - \gamma - \pi_1 - \pi_2 - \pi_3 + \tau - \lambda$	
A → B	Pre-treatment	$\mu + \gamma + \pi_1$	$\mu + \gamma + \pi_3 + \theta$	(d) Kenward and Jones model
	Post-treatment	$\mu + \gamma + \pi_2 + \tau$	$\mu + \gamma - \pi_1 - \pi_2 - \pi_3 - \tau + \lambda$	
B → A	Pre-treatment	$\mu - \gamma + \pi_1$	$\mu - \gamma + \pi_3 - \theta$	
	Post-treatment	$\mu - \gamma + \pi_2 - \tau$	$\mu - \gamma - \pi_1 - \pi_2 - \pi_3 + \tau - \lambda$	

μ	overall mean	π	period effect
τ	treatment effect	λ	carryover effect
θ	residual effect	γ	sequence effect

Model (a) in Table 4.1, due to Willan and Pater [2], assumes that certain components remain constant across phases. In particular, the run-in period is constrained to have the same period effect as the first treatment period, and carryover is assumed identical across the washout and second treatment phases. These constraints have been criticized as unrealistic in clinical applications where phase-specific effects may differ over time [2]. Willan and Pater compared two analyses: post-treatment outcomes alone versus change-from-baseline. They derived estimators for the treatment contrast and its variance under both methods. Under their assumptions, the change-from-baseline approach eliminates the carryover term but can reduce precision (larger mean squared error) relative to post-only analysis.

Model (b) in Table 4.1, proposed by Varma and Chilton [110], extends model (a) by introducing a residual effect θ to capture post-treatment influences at the start of period 2, while retaining a single period effect π . As in model (a), equating the run-in and first treatment periods can be an oversimplification. Patel [113] examined an ANCOVA approach

based on this structure, using baselines to conduct preliminary tests before including them as covariates. His variance calculations were obtained under independence of random errors across periods, an assumption at odds with the correlation structure implied by crossover designs.

Chi [112] proposed the cell-mean model shown in Table 4.1(c). Unlike models (a) and (b), it does not assume that the run-in and first treatment periods share the same period effect, nor that the washout and second treatment periods are equivalent. However, as in model (a), carryover is taken to be constant between the washout and second treatment phases. Chi also allowed a more flexible error structure that can include autoregressive components to account for temporal correlation.

The model adopted in this study is model (d) in Table 4.1 that is proposed by Kenward and Jones [111]. In this model no assumptions are made about the equality of run-in and first period effects, or of wash-out and second period effects; not it is it assumed that the same carryover effect persists from teh was-out into second period. A sequence parameter γ is included; as noted by Kenward and Jones [111], this term should be null under proper randomization, but retaining it ensures that the principal estimators can be written as within-subject contrasts and provides a diagnostic for potential randomization imbalance.

4.2.2 Multivariate Normal Framework for Longitudinal Data

Following Grieve [96], we model the repeated measurements for each participant using a multivariate normal (MVN) distribution with a CS (Compound Symmetry) within-subject covariance. The mean vector stacks the expected responses at the four occasions, and the covariance matrix encodes a common variance on the diagonal and a common covariance off

the diagonal. Let:

$$\begin{aligned}
i &\in \{\text{AB}, \text{BA}\} && (\text{randomization sequence}), \\
j &= 1, \dots, n_i && (\text{subject index within sequence } i), \\
t &= 1, 2, 3, 4, && (\text{measurement occasions per subject}), \\
N &= n_{AB} + n_{BA}, && (\text{total sample size}).
\end{aligned} \tag{4.1}$$

For subject j in sequence i , define the 4-vector of responses $\mathbf{y}_{ij}$ following a multivariate normal (MVN) distribution

$$\mathbf{y}_{ij} = (y_{ij1}, y_{ij2}, y_{ij3}, y_{ij4})^\top \in \mathbb{R}^4, \quad \mathbf{y}_{ij} \sim \mathcal{N}_4(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}), \tag{4.2}$$

with sequence-specific mean vector

$$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})^\top, \tag{4.3}$$

so that μ_{it} is the expected response at occasion t for sequence i . Covariance matrix $\boldsymbol{\Lambda}$ is CS that takes the explicit form

$$\boldsymbol{\Lambda} = \sigma^2[(1 - \rho)\mathbf{I}_4 + \rho \mathbf{1}_4 \mathbf{1}_4^\top] = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \quad \sigma^2 > 0, \quad -\frac{1}{3} < \rho < 1. \tag{4.4}$$

The diagonal entries of $\boldsymbol{\Lambda}$ are all σ^2 (the common marginal variance) and every off-diagonal entry is $\sigma^2 \rho$ (the common covariance), reflecting equal correlation among the four within-subject occasions. Positive definiteness of $\boldsymbol{\Lambda}$ requires $-1/3 < \rho < 1$. The determinant and

inverse, used later for likelihood and posterior derivations, are

$$\det(\mathbf{\Lambda}) = \sigma^8 (1 - \rho)^3 (1 + 3\rho), \quad (4.5)$$

$$\mathbf{\Lambda}^{-1} = \frac{1}{\sigma^2(1 - \rho)(1 + 3\rho)} \begin{bmatrix} 1 + 2\rho & -\rho & -\rho & -\rho \\ -\rho & 1 + 2\rho & -\rho & -\rho \\ -\rho & -\rho & 1 + 2\rho & -\rho \\ -\rho & -\rho & -\rho & 1 + 2\rho \end{bmatrix}. \quad (4.6)$$

4.2.3 Prior distribution

In Bayesian analysis, priors encode pre-data beliefs about unknown parameters. Following Geisser [114] and adopting the CS covariance used in this chapter, we place a weakly informative (improper) prior on the mean components and CS parameters $(\boldsymbol{\mu}, \sigma^2, \rho)$:

$$\pi(\boldsymbol{\mu}, \sigma^2, \rho) \propto \frac{1}{\sigma^2(1 - \rho)(1 + 3\rho)}, \quad \sigma^2 > 0, \quad -\frac{1}{3} < \rho < 1, \quad (4.7)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_{\text{AB}}^\top, \boldsymbol{\mu}_{\text{BA}}^\top)^\top \in \mathbb{R}^8$ collects the four occasion-specific means for each sequence. The factors $(1 - \rho)$ and $(1 + 3\rho)$ in the denominator down-weight boundary values of ρ at which the CS covariance becomes ill-conditioned or singular. Hence (4.7) acts as a gentle guard against values of ρ that would make $\det(\mathbf{\Lambda})$ close to zero, stabilizing computation without materially constraining posterior learning in the interior of the parameter space. Although (4.7) is improper, the resulting posterior is proper under the standard regularity for the MVN/CS model: $\sigma^2 > 0$, $-\frac{1}{3} < \rho < 1$, and a nondegenerate design with adequate replication. In practice, having at least two subjects per sequence (so $N - 2 > 0$) suffices for the variance-component posteriors derived later to be well-defined.

4.2.4 Likelihood function

In Bayesian analysis, the likelihood quantifies how plausible different parameter values are given the observed data; it is the bridge between the statistical model and empirical evidence [115]. In our MVN setup with CS, the likelihood depends on the mean vectors (one per sequence) and the common within-subject covariance $\mathbf{\Lambda}$. The determinant $|\mathbf{\Lambda}|$ reflects overall dispersion, while $\mathbf{\Lambda}^{-1}$ weights within-subject deviations; together they control the shape of the MVN density.

For readability and to match the design of this study, we present the likelihood directly for four measurement occasions per subject: baseline and post-treatment in each of two periods. The single-subject MVN density (with the normalizing constant retained) is

$$f(\mathbf{y}_{ij} \mid \boldsymbol{\mu}_i, \mathbf{\Lambda}) = \frac{1}{(2\pi)^2 |\mathbf{\Lambda}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}_{ij} - \boldsymbol{\mu}_i)^\top \mathbf{\Lambda}^{-1}(\mathbf{y}_{ij} - \boldsymbol{\mu}_i)\right\}. \quad (4.8)$$

For a sample of size n_i in sequence i , the likelihood is the product over j :

$$\begin{aligned} L_i(\boldsymbol{\mu}_i, \mathbf{\Lambda}) &= \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} \mid \boldsymbol{\mu}_i, \mathbf{\Lambda}) \\ &= \frac{1}{(2\pi)^{2n_i} |\mathbf{\Lambda}|^{n_i/2}} \exp\left\{-\frac{1}{2} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \boldsymbol{\mu}_i)^\top \mathbf{\Lambda}^{-1}(\mathbf{y}_{ij} - \boldsymbol{\mu}_i)\right\}. \end{aligned} \quad (4.9)$$

Introduce the sequence mean vector and the centered sum of squares and cross-products (SSCP) matrix

$$\bar{\mathbf{y}}_{i\cdot} := \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij} = (\bar{y}_{i\cdot 1}, \bar{y}_{i\cdot 2}, \bar{y}_{i\cdot 3}, \bar{y}_{i\cdot 4})^\top, \quad (4.10)$$

$$\mathbf{B}_i := \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^\top, \quad (4.11)$$

where $\bar{\mathbf{y}}_{i\cdot}$ summarizes the average four-occasion profile in sequence i , and $\mathbf{B}_i$ measures within-

sequence dispersion around that profile. Using completion of squares,

$$\begin{aligned}
\sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}^{-1} (\mathbf{y}_{ij} - \boldsymbol{\mu}_i) &= \sum_{j=1}^{n_i} [(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) + (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)]^\top \boldsymbol{\Lambda}^{-1} [(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) + (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)] \\
&= \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^\top \boldsymbol{\Lambda}^{-1} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) + n_i (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}^{-1} (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i),
\end{aligned} \tag{4.12}$$

since $\sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) = \mathbf{0}$ makes the cross-term vanish. The first sum can be written using the trace identity

$$\sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot})^\top \boldsymbol{\Lambda}^{-1} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i\cdot}) = \text{tr}(\boldsymbol{\Lambda}^{-1} \mathbf{B}_i). \tag{4.13}$$

Substituting (4.12) and (4.13) into (4.9) yields the sequence-level likelihood in decomposed form:

$$L_i(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}) = \frac{1}{(2\pi)^{2n_i} |\boldsymbol{\Lambda}|^{n_i/2}} \exp\left\{-\frac{1}{2} [n_i (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}^{-1} (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i) + \text{tr}(\boldsymbol{\Lambda}^{-1} \mathbf{B}_i)]\right\}. \tag{4.14}$$

Finally, combining the two sequences with $N = n_{AB} + n_{BA}$ gives the full likelihood

$$\begin{aligned}
L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2, \rho) &= \prod_{i \in \{AB, BA\}} L_i(\boldsymbol{\mu}_i, \boldsymbol{\Lambda}) \\
&= \frac{1}{(2\pi)^{2N} |\boldsymbol{\Lambda}|^{N/2}} \exp\left\{-\frac{1}{2} \sum_{i \in \{AB, BA\}} [n_i (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Lambda}^{-1} (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i) + \text{tr}(\boldsymbol{\Lambda}^{-1} \mathbf{B}_i)]\right\},
\end{aligned} \tag{4.15}$$

with $|\boldsymbol{\Lambda}|$ and $\boldsymbol{\Lambda}^{-1}$ as in (4.5) and (4.6).

Note here that $\boldsymbol{\mu}_i$ stacks the expected values at the four occasions (baseline/post in each period) for sequence i ; $\bar{\mathbf{y}}_{i\cdot}$ is its sample counterpart. The likelihood depends on the data through the sequence means $\bar{\mathbf{y}}_{i\cdot}$ and the SSCP matrices $\mathbf{B}_i$, which neatly separate "between-mean" and "within-sequence" information. It is also worth mentioning that the determinant

factor $|\mathbf{\Lambda}|^{-n/2}$ shrinks the overall scale as $\rho \rightarrow 1$ or $\rho \rightarrow -\frac{1}{3}$, signaling near-singularity. The inverse $\mathbf{\Lambda}^{-1}$ weights deviations; directions that are more variable under $\mathbf{\Lambda}$ receive less weight.

4.2.5 Posterior Distribution

Once the prior distribution is specified and the current data are available, they are combined using Bayes' theorem to produce the posterior distribution. This posterior distribution represents an updated understanding of the parameters, reflecting both the prior information and the observed data. In this context, Bayes' theorem provides a formal mechanism for updating prior beliefs about the population mean, variance, and other characteristics of the distribution based on evidence in the data [115]. In what follows we assume Y represents the observations of the study.

Combining the likelihood (4.15) and the prior (4.7) gives the joint posterior kernel for $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2, \rho)$:

$$\pi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \sigma^2, \rho \mid Y) \tag{4.16}$$

$$\propto \sigma^{-4N-2} (1 - \rho)^{-\frac{3N}{2}-1} (1 + 3\rho)^{-\frac{N}{2}-1} \exp \left\{ -\frac{1}{2} \sum_i [n_i (\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i)^\top \mathbf{\Lambda}^{-1} (\bar{\mathbf{y}}_i - \boldsymbol{\mu}_i) + \text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i)] \right\}. \tag{4.17}$$

Integrating (4.17) over $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ yields the marginal posterior for (σ^2, ρ) :

$$\pi(\sigma^2, \rho \mid Y) \propto |\mathbf{\Lambda}|^{-(N-2)/2} \frac{1}{\sigma^2(1 - \rho)(1 + 3\rho)} \exp \left\{ -\frac{1}{2} \sum_i \text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i) \right\} \tag{4.18}$$

$$\propto (\sigma^2)^{-\{2(N-2)\}-1} (1 - \rho)^{-\frac{3(N-2)}{2}-1} (1 + 3\rho)^{-\frac{(N-2)}{2}-1} \exp \left\{ -\frac{1}{2} \sum_i \text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i) \right\}. \tag{4.19}$$

Using (4.6), the trace term expands as

$$\text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i) = \frac{(1+2\rho) S_i - \rho R_i}{\sigma^2(1-\rho)(1+3\rho)}, \quad S_i := \sum_{t=1}^4 (\mathbf{B}_i)_{tt}, \quad R_i := \sum_{t \neq s} (\mathbf{B}_i)_{ts}. \quad (4.20)$$

The quantities S_i and R_i decompose the trace $\text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i)$ into diagonal (variance) and off-diagonal (covariance) contributions, respectively, matching the CS structure.

Conditional on (σ^2, ρ) , the mean vectors have a multivariate normal posterior:

$$\begin{aligned} \pi(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \mid \sigma^2, \rho, Y) &\propto \exp \left\{ -\frac{1}{2} \sum_i n_i (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i)^\top \mathbf{\Lambda}^{-1} (\bar{\mathbf{y}}_{i\cdot} - \boldsymbol{\mu}_i) \right\} \\ &= \mathcal{N}_8 \left(\begin{bmatrix} \bar{\mathbf{y}}_{AB\cdot} \\ \bar{\mathbf{y}}_{BA\cdot} \end{bmatrix}, \text{diag} \left(\frac{1}{n_{AB}} \mathbf{\Lambda}, \frac{1}{n_{BA}} \mathbf{\Lambda} \right) \right). \end{aligned} \quad (4.21)$$

Equivalently, $\boldsymbol{\mu}_i \mid (\sigma^2, \rho, Y) \sim \mathcal{N}_4(\bar{\mathbf{y}}_{i\cdot}, \mathbf{\Lambda}/n_i)$, independently over i . For further simplification, reparameterize the CS covariance with the variance components

$$\sigma_1^2 = \sigma^2(1+3\rho), \quad \sigma_2^2 = \sigma^2(1-\rho). \quad (4.22)$$

The inverse mapping is $\sigma^2 = \frac{\sigma_1^2 + 3\sigma_2^2}{4}$ and $\rho = \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + 3\sigma_2^2}$, with Jacobian

$$\left| \frac{\partial(\sigma^2, \rho)}{\partial(\sigma_1^2, \sigma_2^2)} \right| = \frac{1}{\sigma_1^2 + 3\sigma_2^2}. \quad (4.23)$$

Define

$$SS_1 = \sum_i \frac{S_i + R_i}{4}, \quad SS_2 = \sum_i \frac{3S_i - R_i}{4}. \quad (4.24)$$

Then the exponent term in (4.19) becomes

$$\sum_i \text{tr}(\mathbf{\Lambda}^{-1} \mathbf{B}_i) = \frac{SS_1}{\sigma_1^2} + \frac{SS_2}{\sigma_2^2}, \quad (4.25)$$

after substituting (4.22) into (4.20). Substituting (4.22) and (4.25) and (4.23) into (4.19) gives the posterior kernel (up to a constant) in (σ_1^2, σ_2^2) :

$$\pi(\sigma_1^2, \sigma_2^2 | Y) \propto (\sigma_1^2)^{-\frac{(N-2)}{2}-1} \exp\left(-\frac{SS_1}{2\sigma_1^2}\right) \cdot (\sigma_2^2)^{-\frac{3(N-2)}{2}-1} \exp\left(-\frac{SS_2}{2\sigma_2^2}\right) \cdot \frac{1}{\sigma_1^2 + 3\sigma_2^2}. \quad (4.26)$$

The factor $1/(\sigma_1^2 + 3\sigma_2^2)$ arises from the Jacobian (4.23). If desired, one may work with the simplified kernel that drops this factor; doing so corresponds to using (σ_1^2, σ_2^2) as primitive parameters with an induced (improper) prior proportional to the Jacobian.

The analysis may impose the prior constraint $\rho > 0$, which is equivalent to requiring the correlated variance to exceed the independent variance, i.e., $\sigma_1^2 > \sigma_2^2$. Enforcing this in the posterior gives

$$\pi(\sigma_1^2, \sigma_2^2 | Y, \sigma_1^2 > \sigma_2^2) = \begin{cases} \frac{\pi(\sigma_1^2, \sigma_2^2 | Y)}{\mathbb{P}(\sigma_1^2 > \sigma_2^2 | Y)}, & \sigma_1^2 > \sigma_2^2, \\ 0, & \text{otherwise.} \end{cases} \quad (4.27)$$

To obtain a convenient pivotal ratio, define

$$\phi = \frac{1}{3} \cdot \frac{\sigma_1^2}{\sigma_2^2} \cdot \frac{SS_2}{SS_1}, \quad \psi = \sigma_2^2, \quad (4.28)$$

with Jacobian

$$J = \left| \frac{\partial(\sigma_1^2, \sigma_2^2)}{\partial(\phi, \psi)} \right| = 3\psi \frac{SS_1}{SS_2}. \quad (4.29)$$

Substituting (4.28) and (4.29) into the (σ_1^2, σ_2^2) posterior (from the previous step) yields, up to a constant,

$$\pi(\psi, \phi | Y) \propto \psi^{-2N+3} \phi^{-\frac{(N-2)}{2}-1} \exp\left\{-\frac{SS_1}{2\psi} \left(1 + \frac{1}{3\phi}\right)\right\}. \quad (4.30)$$

Integrating out ψ gives the marginal density of ϕ :

$$\pi(\phi | Y) \propto \phi^{-\frac{(N-2)}{2}-1} \left(1 + \frac{1}{3\phi}\right)^{-2(N-2)}. \quad (4.31)$$

Equivalently, recognizing the standard ratio form,

$$\phi \sim F_{3(N-2), N-2}. \quad (4.32)$$

The posterior probability of the constraint $\sigma_1^2 > \sigma_2^2$ can be expressed via ϕ . Since

$$\sigma_1^2 > \sigma_2^2 \iff \phi > \frac{1}{3} \frac{SS_2}{SS_1},$$

we obtain the upper-tail F probability

$$\mathbb{P}(\sigma_1^2 > \sigma_2^2 | Y) = \mathbb{P}\left(\phi > \frac{1}{3} \frac{SS_2}{SS_1}\right) = 1 - F_{3(N-2), N-2}\left(\frac{1}{3} \frac{SS_2}{SS_1}\right), \quad (4.33)$$

where $F_{\nu_1, \nu_2}(\cdot)$ denotes the CDF of the F -distribution with ν_1, ν_2 degrees of freedom.

The Kenward–Jones model (with residual/baseline offset θ) can be written in cell-

mean form as

$$\begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{14} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{24} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & -1 & -1 & -1 & 0 & 1 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & -1 & 0 \\ 1 & -1 & -1 & -1 & -1 & 1 & 0 & -1 \end{pmatrix}}_{\mathbb{X} \ (8 \times 8)} \begin{pmatrix} \mu \\ \gamma \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \tau \\ \theta \\ \lambda \end{pmatrix}.$$

In this formulation, each row of the design matrix corresponds to one of the eight cell means μ_{ij} , and each column represents a specific model parameter. This matrix representation is not only mathematically efficient but also theoretically consistent. It ensures that the relationship between the cell means and the model parameters is clearly defined, making it possible to understand how each parameter contributes to the observed outcomes. Each cell mean is expressed as a specific combination of these parameters, reflecting the logical structure of the crossover design. For instance, the cell means for the first sequence (AB) are influenced by the overall mean, the group effect, the period effects, and the treatment effect, while the means for the second sequence (BA) are adjusted by the opposite signs of the group and treatment effects, consistent with the reversed sequence. To estimate the model parameters, define the between-sequence differences at each occasion,

$$d_t = \mu_{1t} - \mu_{2t}, \quad t = 1, 2, 3, 4,$$

so under the model

$$d_1 = 2\gamma, \quad d_2 = 2\gamma + 2\tau, \quad d_3 = 2\gamma + 2\theta, \quad d_4 = 2\gamma - 2\tau + 2\lambda.$$

Hence the non-iterative estimators are

$$\hat{\gamma} = \frac{d_1}{2} = \frac{\bar{y}_{1.1} - \bar{y}_{2.1}}{2}, \quad (4.34)$$

$$\hat{\tau} = \frac{d_2 - d_1}{2} = \frac{-\bar{y}_{1.1} + \bar{y}_{1.2} + \bar{y}_{2.1} - \bar{y}_{2.2}}{2}, \quad (4.35)$$

$$\hat{\theta} = \frac{d_3 - d_1}{2} = \frac{-\bar{y}_{1.1} + \bar{y}_{1.3} + \bar{y}_{2.1} - \bar{y}_{2.3}}{2}, \quad (4.36)$$

$$\hat{\lambda} = \frac{d_4 - 2\hat{\gamma} + 2\hat{\tau}}{2} = \frac{-2\bar{y}_{1.1} + \bar{y}_{1.2} + \bar{y}_{1.4} + 2\bar{y}_{2.1} - \bar{y}_{2.2} - \bar{y}_{2.4}}{2}. \quad (4.37)$$

These are linear in the eight means and provide a convenient right-inverse on the estimable subspace, without requiring a full 8×8 inverse. Stack the eight cell means as $\boldsymbol{\mu}_{\text{cell}} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{14}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})^\top$ and write

$$\boldsymbol{d} = (d_1, d_2, d_3, d_4)^\top = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}}_{\mathbb{D} \ (4 \times 8)} \boldsymbol{\mu}_{\text{cell}},$$

and

$$\begin{pmatrix} \hat{\gamma} \\ \hat{\tau} \\ \hat{\theta} \\ \hat{\lambda} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & 0 \\ -1 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}}_{\boldsymbol{T} \ (4 \times 4)} \boldsymbol{d} = \underbrace{(\boldsymbol{T}\mathbb{D})}_{K \ (4 \times 8)} \boldsymbol{\mu}_{\text{cell}}.$$

Thus K is a correct linear map from the eight means to $(\gamma, \tau, \theta, \lambda)^\top$. To find the posterior distribution of parameters, let $q = \frac{1}{n_{AB}} + \frac{1}{n_{BA}}$. Under compound symmetry $\boldsymbol{\Lambda} = \sigma^2[(1 -$

$\rho)\mathbf{I}_4 + \rho \mathbf{1}_4 \mathbf{1}_4^\top]$ with $\sigma_2^2 = \sigma^2(1 - \rho)$, we have

$$\mathbf{d} = (\bar{\mathbf{y}}_{\text{AB}\cdot} - \bar{\mathbf{y}}_{\text{BA}\cdot}) \sim \mathcal{N}_4(\mathbb{E}\mathbf{d}, q\mathbf{\Lambda}), \quad \begin{pmatrix} \hat{\gamma} \\ \hat{\tau} \\ \hat{\theta} \\ \hat{\lambda} \end{pmatrix} = \mathbf{T} \mathbf{d}.$$

Hence, conditional on (σ^2, ρ) ,

$$\begin{pmatrix} \gamma \\ \tau \\ \theta \\ \lambda \end{pmatrix} \Big| \sigma^2, \rho, Y \sim \mathcal{N}_4 \left(\begin{pmatrix} \hat{\gamma} \\ \hat{\tau} \\ \hat{\theta} \\ \hat{\lambda} \end{pmatrix}, \mathbf{V} \right), \quad \mathbf{V} = q\mathbf{T}\mathbf{\Lambda}\mathbf{T}^\top. \quad (4.38)$$

The diagonal entries simplify to

$$\text{Var}(\gamma \mid \sigma^2, \rho) = \frac{q\sigma^2}{4}, \quad \text{Var}(\tau \mid \sigma^2, \rho) = \frac{q\sigma_2^2}{2}, \quad \text{Var}(\theta \mid \sigma^2, \rho) = \frac{q\sigma_2^2}{2}, \quad \text{Var}(\lambda \mid \sigma^2, \rho) = \frac{3q\sigma_2^2}{2},$$

with off-diagonals given by $\mathbf{V} = q\mathbf{T}\mathbf{\Lambda}\mathbf{T}^\top$, where,

$$\sigma_1^2 = \sigma^2(1 + 3\rho), \quad \sigma_2^2 = \sigma^2(1 - \rho), \quad \sigma^2 = \frac{\sigma_1^2 + 3\sigma_2^2}{4}.$$

To derive the marginal posterior distributions of τ, θ , and λ , we first need the posterior for σ_2^2 . Using the (σ_1^2, σ_2^2) reparametrization and the simplified kernel in (4.26) that factorizes in (σ_1^2, σ_2^2) , we obtain

$$\pi(\sigma_2^2 \mid Y) \propto (\sigma_2^2)^{-\frac{3N}{2}+2} \exp\left(-\frac{SS_2}{2\sigma_2^2}\right). \quad (4.39)$$

Thus $\sigma_2^2 \mid Y$ is inverse-gamma with shape $\alpha = \frac{3(N-2)}{2}$ and scale $\beta = \frac{SS_2}{2}$ in the convention

$\pi(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2)$, a distribution commonly used for variance parameters in Bayesian analysis. The exponent $-3N/2 + 2$ is determined by the data size N , while the exponential decay term $\exp\left(-\frac{SS_2}{2\sigma_2^2}\right)$ is directly scaled by the sum of squares SS_2 , which captures the variability in the observed data. This distribution effectively expresses the uncertainty about σ_2^2 given the data.

Conditional on (σ^2, ρ) , we have (from the Kenward–Jones contrasts; see (4.38))

$$\tau \mid \sigma_2^2, Y \sim \mathcal{N}\left(\hat{\tau}, \frac{q\sigma_2^2}{2}\right), \quad \theta \mid \sigma_2^2, Y \sim \mathcal{N}\left(\hat{\theta}, \frac{q\sigma_2^2}{2}\right), \quad \lambda \mid \sigma_2^2, Y \sim \mathcal{N}\left(\hat{\lambda}, \frac{3q\sigma_2^2}{2}\right). \quad (4.40)$$

Having established the posterior distribution of σ_2^2 , the next objective is to derive the marginal posterior distributions of τ , θ and λ . This is achieved by integrating σ_2^2 out of the joint posterior distribution of these parameters in (4.38), ensuring that the resulting distributions fully account for the uncertainty in σ_2^2 . The process begins by expressing the joint distribution of each parameter with σ_2^2 and then performing the integration. As we see below, the resulting posterior distributions are the familiar “shifted-and-scaled Student- t ” posteriors with $\nu = 2\alpha = 3(N - 2)$ degrees of freedom, obtained by mixing the normal conditionals (4.40) with the inverse-gamma for σ_2^2 in (4.39). Each density is centered at its posterior mean $(\hat{\tau}, \hat{\theta}, \hat{\lambda})$; the scale reflects both sampling variability (through q) and the uncertainty in σ_2^2 (through SS_2 and α). As N grows, $\nu = 3(N - 2)$ increases and the posteriors become increasingly concentrated. For τ , this integration produces the following marginal posterior distribution:

$$\pi(\tau|Y) \propto \left[SS_2 + \frac{2}{q}(\tau - \hat{\tau})^2\right]^{-\frac{3N-5}{2}} \propto \left[\frac{SS_2}{2} + \frac{(\tau - \hat{\tau})^2}{q}\right]^{-\left(\alpha + \frac{1}{2}\right)}, \quad (4.41)$$

which is a shifted and scaled t -distribution, centered around the posterior mean $\hat{\tau}$, i.e.,

$$\tau \mid Y \sim t_\nu\left(\hat{\tau}, \sqrt{\frac{q}{2} \cdot \frac{SS_2}{\alpha}}\right),$$

with $\nu = 2\alpha = 3(N - 2)$ degrees of freedom. The quadratic term $(\tau - \hat{\tau})^2$ reflects how the probability density decreases as τ deviates from its posterior mean. The scaling term $SS^2 + \frac{2}{q}(\tau - \hat{\tau})^2$ ensures that the distribution properly accounts for the variability in the data. The marginal posterior distribution of θ is obtained in a similar way:

$$\pi(\theta|Y) \propto \left[SS_2 + \frac{2}{q}(\theta - \hat{\theta})^2 \right]^{-\frac{3N-5}{2}} \quad \text{or} \quad \theta | Y \sim t_\nu \left(\hat{\theta}, \sqrt{\frac{q}{2} \cdot \frac{SS_2}{\alpha}} \right). \quad (4.42)$$

This distribution shares the same structure, with the central value set at $\hat{\theta}$, the posterior mean of θ . The quadratic term ensures that the posterior is concentrated around this mean, while the scaling term adjusts the variability according to the sum of squares SS_2 and the data. For λ the marginal posterior is defined as:

$$\pi(\lambda|Y) \propto \left[SS_2 + \frac{2}{3q}(\lambda - \hat{\lambda})^2 \right]^{-\frac{3N-5}{2}} \quad \text{or} \quad \lambda | Y \sim t_\nu \left(\hat{\lambda}, \sqrt{\frac{3q}{2} \cdot \frac{SS_2}{\alpha}} \right). \quad (4.43)$$

These results demonstrate that the marginal posterior distributions of τ , θ , and λ are shifted and scaled t-distributions. This form directly results from integrating out σ_2^2 , accounting for its uncertainty. Each distribution is centered at its posterior mean $(\hat{\tau}, \hat{\theta}, \hat{\lambda})$, with the scaling terms $SS_2 + \frac{2}{q}(\text{parameter} - \widehat{\text{parameter}})^2$ reflecting the data variability. The degrees of freedom $-\frac{(3N-5)}{2}$ determine the shape, becoming sharper with larger datasets. In Grieve's work [116], two initial models, M_0 and M_1 , were introduced, allowing for the estimation of the treatment effect τ . Subsequently, Grieve expanded this approach by proposing four models, labeled M_2 , M_{11} , M_{12} , and M_0 [96]. These models provide a broader framework for estimating treatment effects under varying assumptions. They are defined as follows:

Table 4.2: Conditional posterior distributions in model $M_i : N(\alpha, \beta m \sigma^2)$

Model	Parameter	Posterior Mean (α)	Posterior Scaling (β)
$M_2 : \tau, \theta, \lambda(\text{unrestricted})$	τ	$\hat{\tau}$	$\frac{1}{2}$
	θ	$\hat{\theta}$	$\frac{1}{2}$
	λ	$\hat{\lambda}$	$\frac{3}{2}$
$M_{11} : \tau, \lambda(\theta = 0)$	τ	$\hat{\tau} - \frac{\hat{\theta}}{2}$	$\frac{3}{8}$
	λ	$\hat{\lambda} - \hat{\theta}$	1
$M_{12} : \tau, \theta(\lambda = 0)$	τ	$\hat{\tau} - \frac{\hat{\lambda}}{2}$	$\frac{1}{8}$
	θ	$\hat{\theta} - \frac{\hat{\lambda}}{3}$	$\frac{1}{3}$
$M_0 : \tau(\theta = 0, \lambda = 0)$	τ	$\hat{\tau} - \frac{\hat{\lambda}}{2}$	$\frac{1}{8}$

Table 2 provides a summary of the conditional posterior distributions for the parameters of interest across the four models: M_2 , M_{11} , M_{12} and M_0 .

M_2 is the most general model, where both θ and λ are freely estimated, capturing any potential interaction between treatments and any carryover effects from one period to the next.

M_{11} is a restricted model where $\theta = 0$, assuming no interaction effect between treatments, but still allowing λ to account for potential carryover effects.

M_{12} is another restricted model where $\lambda = 0$, eliminating any carryover effects, but allowing θ to account for potential interaction effects.

M_0 is the most constrained model, assuming both $\theta = 0$ and $\lambda = 0$, effectively treating the treatment effect τ as independent of any interaction or carryover influences.

By systematically applying these models, the analysis can explore how different assumptions about θ and λ affect the estimation of τ . The conditional posterior distributions of τ under each model are derived directly using the multivariate normal structure of (11):

In M_{11} , where $\theta = 0$, the posterior distribution of τ is given by:

$$\pi(\tau|\theta = 0, \sigma_1^2, \sigma_2^2, Y) \sim \mathcal{N}(\hat{\tau} - \frac{\hat{\theta}}{2}, \frac{3q\sigma_2^2}{8}) \quad (4.44)$$

In M_{12} , where $\lambda = 0$, the conditional posterior of τ is:

$$\pi(\tau|\lambda = 0, \sigma_1^2, \sigma_2^2, Y) \sim \mathcal{N}(\hat{\tau} - \frac{\hat{\lambda}}{2}, \frac{q\sigma_2^2}{8}) \quad (4.45)$$

In M_0 , where $\theta = 0$ and $\lambda = 0$, the posterior distribution simplifies to:

$$\pi(\tau|\theta = 0, \lambda = 0, \sigma_1^2, \sigma_2^2, Y) \sim \mathcal{N}(\hat{\tau} - \frac{\hat{\lambda}}{2}, \frac{q\sigma_2^2}{8}) \quad (4.46)$$

These conditional posterior distributions illustrate how the estimation of τ is directly influenced by the assumptions about θ and λ . To further refine the analysis of the model parameters, the focus shifts to deriving the marginal posterior distributions of σ_2^2 under each of the specified models. This step is essential because the marginal distribution of σ_2^2 directly influences the conditional posterior distributions of the treatment effect τ , the interaction effect θ , and the carryover effect λ . The approach begins with the posterior distribution of σ_2^2 derived in (12), which is an inverse-gamma-like distribution characterized by its dependence on SS_2 , the sum of squares, and the model-specific parameters.

Starting from the general form in (12), the marginal posterior distribution of σ_2^2 is adjusted for each model, reflecting the influence of the model-specific assumptions about θ and λ .

For M_{11} , where $\theta = 0$, the posterior distribution is given by:

$$\pi(\sigma_2^2|Y, M_{11}) \propto (\sigma_2^2)^{-3N/2+3/2} \exp(-\frac{SS_2 + 2\hat{\theta}^2/q}{2\sigma_2^2})$$

For M_{12} , where $\lambda = 0$, the marginal distribution becomes:

$$\pi(\sigma_2^2|Y, M_{12}) \propto (\sigma_2^2)^{-3N/2+3/2} \exp\left(-\frac{SS_2 + 2\hat{\lambda}^2/(3q)}{2\sigma_2^2}\right)$$

In M_0 , where both $\theta = 0$ and $\lambda = 0$, the distribution is further simplified:

$$\pi(\sigma_2^2|Y, M_0) \propto (\sigma_2^2)^{-3N/2+2/2} \exp\left(-\frac{SS_2 + 3\hat{\theta}^2/q - 2\hat{\theta}\hat{\lambda}/q + \hat{\lambda}^2/q}{2\sigma_2^2}\right)$$

These expressions represent scaled inverse-gamma distributions, where the shape and scale parameters are influenced by the model-specific terms $\hat{\theta}$ and $\hat{\lambda}$. These terms directly capture the impact of the interaction and carryover effects on the variance component σ_2^2 . The exponent $-3N/2 + k/2$ reflects the degrees of freedom adjusted for the model complexity, while the scale parameter in the exponent incorporates the sum of squares SS_2 adjusted by model-specific terms.

Having established the marginal posterior distributions of σ_2^2 for each model, the next step is to derive the marginal posterior distributions of τ , θ and λ by integrating out σ_2^2 . For τ under M_{11} , this integration leads to:

$$\pi(\tau|Y, M_{11}) \propto [SS_2 + \frac{2\hat{\theta}^2}{q} + \frac{8}{3q}(\tau - \hat{\tau} + \frac{\hat{\theta}}{2})^2]^{-\frac{3N-4}{2}} \quad (4.47)$$

Under M_{12} , the posterior of τ is adjusted as follows:

$$\pi(\tau|Y, M_{12}) \propto [SS_2 + \frac{2\hat{\lambda}^2}{3q} + \frac{8}{q}(\tau - \hat{\tau} + \frac{\hat{\lambda}}{2})^2]^{-\frac{3N-4}{2}} \quad (4.48)$$

For M_0 , the most restricted model, the posterior becomes:

$$\pi(\tau|Y, M_0) \propto [SS_2 + \frac{3\hat{\theta}^2}{q} - \frac{2\hat{\theta}\hat{\lambda}}{q} + \frac{\hat{\lambda}^2}{q} + \frac{8}{q}(\tau - \hat{\tau} + \frac{\hat{\theta}}{2})^2]^{-\frac{3N-3}{2}} \quad (4.49)$$

The analysis extends to the posterior distributions of λ and θ . For λ under M_{11} , the posterior

distribution is:

$$\pi(\lambda|Y, M_{11}) \propto [SS_2 + \frac{2\theta^2}{q} + \frac{1}{q}(\lambda - \hat{\lambda} + \hat{\theta}^2)]^{-\frac{3N-4}{2}} \quad (4.50)$$

For θ under M_{12} , the distribution is expressed as:

$$\pi(\theta|Y, M_{12}) \propto [SS_2 + \frac{2\hat{\lambda}^2}{3q} + \frac{3}{q}(\theta - \hat{\theta} + \frac{\hat{\lambda}^2}{3})]^{-\frac{3N-4}{2}} \quad (4.51)$$

These posterior distributions comprehensively describe the uncertainty surrounding the model parameters, conditioned on the data and model assumptions. Each distribution is shaped by the sum of squares SS_2 , scaled by model-specific adjustments involving $\hat{\theta}$ and $\hat{\lambda}$, with quadratic terms reflecting deviations from posterior means.

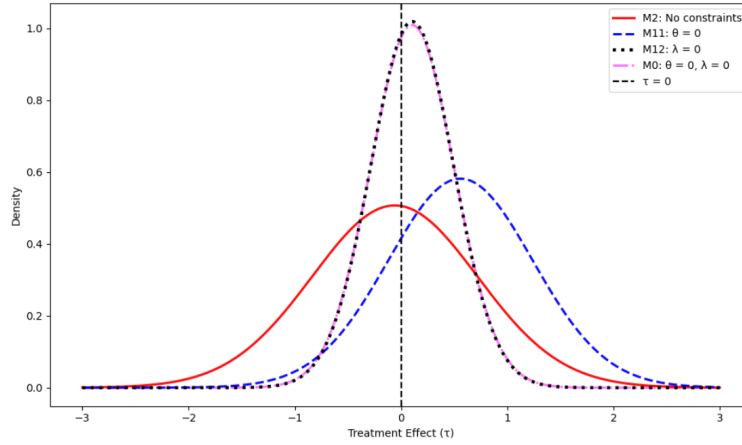


Figure 4.1: Posterior Distribution of Treatment Effect (τ) under four models, each imposing different constraints on the carryover (θ) and interaction (λ) effects

Figure 4.1 displays the posterior distributions of the treatment effect τ under four models, each imposing different constraints on the carryover (θ) and interaction (λ) effects. Peaks closer to zero with narrower spreads (like in M_0 and M_{12}) suggest higher certainty around a null or small effect. The curve labels directly indicate which model each distribution corresponds to.

Table 4.3: Summary statistics for ADAS-Cog data

Parameter	Description	Estimate
$\hat{\gamma}$	Sequence effect	0.051765
$\hat{\tau}$	Treatment effect	-0.059074
$\hat{\theta}$	Period effect	-1.236044
$\hat{\lambda}$	Carryover effect	-0.327132
SS_1	Between-sequence variation	9157.411232
SS_2	Within-sequence variation	1199.005485

The summary statistics obtained from the ADAS-Cog data, as reported in Table 4.3, include estimates of the γ , treatment effect τ , period effect θ , carryover effect λ , and the associated sums of squares SS_1 and SS_2 . These values form the basis for evaluating the posterior distributions of the treatment effect under the various candidate models, as presented in the results that follow.

Table 4.4: Posterior Probability of Positive Treatment Effect Under Each Model

Model	$\mathbb{P}(\tau > 0 \mid \mathbf{Y}, \mathbf{M})$
M_2	0.470059
M_{11}	0.792143
M_{12}	0.605225
M_0	0.604232

In this setting, we are confronted with the challenge of selecting the most appropriate model for making inference about the treatment effect τ . Table 4.4 displays the posterior probabilities that $\tau > 0$ under four competing models, each differing in whether they constrain the carryover effect θ and interaction term λ . The probabilities range from 0.47 under the fully unconstrained model M_2 to 0.79 under model M_{11} , where $\theta = 0$. These differences highlight the sensitivity of inference to modeling assumptions. Specifically, when fewer restrictions are imposed, as in M_2 , the uncertainty about nuisance parameters leads to more conservative conclusions about treatment efficacy. In contrast, constrained models such as M_{11} suggest stronger evidence for a positive treatment effect. This variation underscores the importance of model selection and the potential value of approaches that account for model

uncertainty, such as Bayesian model averaging, to avoid misleading or overly optimistic conclusions.

4.2.6 Bayes Factors

In this section, it is demonstrated that posterior inferences in two-period crossover studies with two baselines are highly sensitive to the choice of model, particularly when assessing the presence of a carryover effect. Traditional approaches, such as Grizzle’s method [6], rely on preliminary hypothesis testing, which has been shown to result in inflated Type I error rates and to overlook model uncertainty by treating the selected model as definitively correct. To overcome these limitations, the use of Bayes factors is proposed as a more robust and probabilistic framework for model comparison, avoiding the need for pre-testing and enabling a more nuanced evaluation of competing models. The concept of the Bayes factor was first introduced by Harold Jeffreys [117]. The Bayes factor provides a quantitative measure of the evidence supporting one statistical model over another [118]. It is formally defined as the ratio of the marginal likelihoods of the data under two competing models:

$$BF_{01} = \frac{\mathbb{P}(\text{Data}|M_0)}{\mathbb{P}(\text{Data}|M_1)}$$

The Bayes factor is a key component of Bayesian hypothesis testing [119],[120]. Unlike frequentist approaches, which rely on p -values to make decisions about rejecting hypotheses, Bayesian hypothesis testing uses Bayes factors as the basis for decision-making [121]. Typically, this involves setting lower and upper thresholds to guide interpretation. A Bayes factor exceeding the upper threshold is considered strong evidence in favor of the alternative hypothesis, while a value below the lower threshold supports the null hypothesis. Values falling between these boundaries indicate that the evidence is inconclusive [122]. To aid interpretation, researchers often use heuristic classification schemes for Bayes factors, one of

which is presented in Table 1 [123].

Table 4.5: Interpretation of Bayes factors in support of M_0 (i.e., BF_{01})

Bayes factor	Evidence category
> 100	Extreme evidence for M_0
$30-100$	Very strong evidence for M_0
$10-30$	Strong evidence for M_0
$3-10$	Moderate evidence for M_0
$1-3$	Anecdotal evidence for M_0
1	No evidence
$1/3-1$	Anecdotal evidence for M_1
$1/10-1/3$	Moderate evidence for M_1
$1/30-1/10$	Strong evidence for M_1
$1/100-1/30$	Very strong evidence for M_1
$< 1/100$	Extreme evidence for M_1

The choice of prior distributions plays a crucial role in Bayesian model comparison, particularly in the calculation of Bayes factors. A known challenge arises when improper priors are used, because Bayes factor computation requires a proper prior, and often a different functional form than what is used for estimating posterior distributions. Spiegelhalter and Smith [124] addressed this issue by introducing an approach based on invariance arguments. Their method focuses on the comparison of two nested linear models with Gaussian error structure, denoted $M_0 \subset M_1$, where the models are formulated as

$$Y \sim \mathcal{N}(A_i \theta_i, \sigma^2 I_N)$$

with Y representing the observed data, A_i the design matrices, θ_i the parameter vectors, and σ^2 the unknown variance.

To compute the Bayes factor comparing M_0 against M_1 , Spiegelhalter and Smith [124] express

it as

$$B_{01} = \frac{\iint p(Y \mid A_0, \theta_0, \sigma^2 I_n) \pi(\theta_0, \sigma^2 \mid A_0) d\theta_0 d\sigma^2}{\iint p(Y \mid A_1, \theta_1, \sigma^2 I_n) \pi(\theta_1, \sigma^2 \mid A_1) d\theta_1 d\sigma^2}$$

Under their formulation, the joint prior on the model parameters (θ_i, σ^2) , conditional on model structure A_i , takes the form of a normal-inverse- $\mathcal{X}^2$ distribution:

$$\pi(\theta_i, \sigma^2 \mid A_i) = c_i (2\pi)^{-p_i/2} (\sigma^2)^{-(p_i/2+1)} \quad (4.52)$$

where c_i is a constant (not integrated out in the improper case) and p_i is the number of parameters under model A_i . Using this setup, the Bayes factor comparing a reduced model M_0 to a more complex model M_1 is derived as:

$$B_{01} = \frac{c_0}{c_1} \left(\frac{|\mathbf{A}_1^\top \mathbf{A}_1|}{|\mathbf{A}_0^\top \mathbf{A}_0|} \right)^{1/2} \left[1 + \frac{(p_1 - p_0)F}{N - p_1} \right]^{-N/2} \quad (4.53)$$

where F is the standard F-statistic for comparing nested models. However, this expression depends on the unknown ratio $\frac{c_0}{c_1}$, rendering the Bayes factor indeterminate without additional assumptions.

To resolve this, Spiegelhalter and Smith introduced the use of an imaginary training sample [125]. This hypothetical dataset is chosen to represent the minimal sample size sufficient to estimate parameters under both models and maximize support for the simpler model M_0 . Under these conditions, the F-statistic reduces to zero, and the ratio of constants simplifies as:

$$\frac{c_0}{c_1} \approx \left(\frac{|\mathbf{E}_1^\top \mathbf{E}_1|}{|\mathbf{E}_0^\top \mathbf{E}_0|} \right)^{-1/2}$$

In contrast, Grieve [116] proposed a more pragmatic solution in the context of two-period crossover trials, where two variance components are of interest: between-patient and within-patient variances. Instead of redefining the prior structure to satisfy Bayes factor requirements, Grieve retained standard improper priors and acknowledged that the prior for the

variance parameters takes the form of

$$\pi(\sigma_1^2, \sigma_2^2) \propto (\sigma_1^2 \sigma_2^2)^{-1}$$

where $\sigma_1^2 = \sigma^2[1 + (k - 1)\rho]$ and $\sigma_2^2 = \sigma^2(1 - \rho)$. Grieve [96] noted that while such a prior is incompatible with the Bayes factor derivation that requires a proper prior, the practical impact is often negligible, affecting the degrees of freedom in the posterior distributions only marginally. Therefore, his analysis favored analytical tractability and adopted the improper prior structure for posterior derivation, without aiming to compute Bayes factors directly.

The key distinction between the two approaches lies in their treatment of prior specification and the trade-offs involved. Spiegelhalter and Smith [124] proposed a method based on proper priors tailored for model comparison via Bayes factors. Their approach involves careful calibration, such as the introduction of an imaginary training sample, to resolve indeterminacies caused by improper priors. While theoretically sound and well-suited for formal model selection, this method introduces added complexity to the posterior analysis and shifts the effective degrees of freedom.

In this analysis, we adopt the approach used by Grieve [96], which employs standard improper priors for the variance components. This choice simplifies the derivation of posterior distributions and avoids the computational overhead of training sample calibration. Although the use of improper priors precludes a formal Bayes factor comparison, the practical impact is minimal, typically resulting in only a small change in the degrees of freedom used in posterior distributions. This reflects a deliberate trade-off: by prioritizing analytical simplicity and interpretability, we accept a modest theoretical limitation in exchange for more direct inference on treatment effects.

In Bayesian analysis of two-period crossover trials with repeated measures, model comparison relies on the marginal likelihood, which is obtained by integrating the likelihood function over the prior distribution of the model's parameters. Each model encodes specific

assumptions, such as the presence or absence of treatment or period effects, and is evaluated using Bayes factors, which quantify the relative evidence in favor of one model while accounting for both fit and complexity.

The general expression for the marginal likelihood of the saturated model M_s , which imposes no constraints and allows full flexibility in estimating all treatment-period mean parameters, is:

$$\mathbb{P}(Y \mid M_s) = b_s w_s \Gamma\left(\frac{N}{2}\right) \left(\frac{SS_1}{2}\right)^{-N/2} \Gamma\left(\frac{N(k-1)}{2}\right) \left(\frac{SS_2}{2}\right)^{-N(k-1)/2} \quad (4.54)$$

This assumption is encoded In this formulation:

- N is the number of subjects,
- k is the number of repeated measures per subject are the between-subject and within-subject sums of squares,
- SS_1 and SS_2 are the between-subject and within-subject sums of squares,
- b_s and w_s are constants depending on the model's prior and design structure.

To illustrate this in the specific case of a two-treatment, two-period crossover design, consider $g = 2$ treatment sequences and $k = 4$ repeated measures per subject. This leads to a total of 8 treatment-period mean parameters. Substituting $g = 2$, $k = 4$ into the general formula yields the saturated model likelihood:

$$\mathbb{P}(Y \mid M_2) = b_2 w_2 \Gamma\left(\frac{N}{2}\right) \left(\frac{SS_1}{2}\right)^{-N/2} \Gamma\left(\frac{3N}{2}\right) \left(\frac{SS_2}{2}\right)^{-3N/2} \quad (4.55)$$

This saturated model serves as a reference point for comparison with more structured models. One such structured model, denoted M_{11} , imposes the assumption of no period effect.

This assumption is encoded through a contrast vector that defines a linear constraint on the treatment-period means. Specifically, the constraint $\frac{1}{2} \begin{pmatrix} -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \end{pmatrix} \boldsymbol{\mu} = 0$. To account for repeated measurements and group structure, the design matrix is built using

a Kronecker product $D = \text{diag}(\frac{1}{n_{AB}}, \frac{1}{n_{BA}}) \otimes \mathbb{I}_4$. Under this constraint, the marginal likelihood for model M_{11} becomes:

$$\mathbb{P}(Y \mid M_{11}) = b_2 w_{11} \Gamma\left(\frac{N}{2}\right) \left(\frac{SS_1}{2}\right)^{-N/2} \Gamma\left(\frac{3N}{2}\right) \left(\frac{2}{q}\right)^{1/2} \left(\frac{SS_2 + \frac{2\hat{\theta}^2}{q}}{2}\right)^{-3N/2} \quad (4.56)$$

Here, $\hat{\theta}$ represents the estimated period effect (which is being constrained to zero in this model), and q is a known scalar derived from the contrast vector and the design. The term involving $\hat{\theta}$ reflects the residual deviation from the hypothesized constraint; it effectively penalizes the likelihood if the observed data strongly contradict the no-period-effect assumption.

From (27) and (28), the Bayes factor comparing M_{11} (the model without period effect) to the saturated model M_2 is given by:

$$B_{11,2} = \frac{\mathbb{P}(Y \mid M_{11})}{\mathbb{P}(Y \mid M_2)} = \frac{w_{11}}{w_2} \left(\frac{2}{q}\right)^{1/2} \left(1 + \frac{2\hat{\theta}^2}{qSS_2}\right)^{-3N/2} \quad (4.57)$$

To evaluate the strength of evidence in favor of the reduced model M_{11} , which assumes no period effect, the Bayes factor $B_{11,2}$ is used to compare it against the saturated model M_2 . As shown earlier, the general form of this Bayes factor includes an undefined constant ratio $\frac{W_{11}}{W_2}$, which arises from integration over improper priors. To resolve this, the approach proposed by Spiegelhalter and Smith is applied, wherein the constant is calibrated by assuming a minimal hypothetical dataset that just allows estimation under both models.

Specifically, under the smallest possible sample size, assuming that the constrained hypothesis is true (i.e., $\hat{\theta} = 0$), the Bayes factor should equal 1, since the reduced model and the saturated model are equally plausible in the absence of any evidence favoring the more complex model. Applying this condition yields the following calibrated constant ratio (See [116] for more details).

$$\frac{w_{11}}{w_2} = \left(\frac{3}{2}\right)^{\frac{1}{2}}.$$

Substituting this into the original expression leads to the fully specified Bayes factor:

$$B_{11,2} = \left(\frac{3}{2q}\right)^{1/2} \left(1 + \frac{2\hat{\theta}^2}{q \cdot SS_2}\right)^{-3N/2} \quad (4.58)$$

This formulation allows the Bayes factor to be computed directly from the observed data without requiring evaluation of arbitrary constants, and appropriately penalizes lack of fit to the no-period-effect hypothesis.

Similarly, a second reduced model M_{12} is introduced based on a different linear contrast:

$$\frac{1}{2} \begin{pmatrix} -2 & 1 & 0 & 1 & 2 & -1 & 0 & -1 \end{pmatrix} \boldsymbol{\mu} = 0$$

which may correspond to an alternative hypothesis, such as the absence of a carryover effect, depending on the parameterization of the means. Applying the same calibration principle and derivation as before, the Bayes factor comparing M_{12} to M_2 is:

$$B_{12,2} = \left(\frac{3}{2q}\right)^{1/2} \left(1 + \frac{2\hat{\lambda}^2}{3qSS_2}\right)^{-3N/2} \quad (4.59)$$

where $\hat{\lambda}$ is the estimated contrast associated with the hypothesis in M_{12} . Following the derivation of Bayes factors for models with individual constraints, attention is now turned to the most restrictive model in the analysis, denoted M_0 . This model assumes that both the period effect (θ) and the carryover effect (λ) are absent. In contrast to the previous models M_{11} and M_{12} , where either the period or the carryover effect was removed, the null model M_0 imposes both constraints simultaneously. This additional structure allows for the assessment of whether a fully simplified model provides a sufficiently good explanation of the data, or whether one or both of these effects are necessary.

To formally define model M_0 , two linear contrast vectors are applied to the full vector of treatment-period means $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \dots, \mu_{24})^\top$. These contrast vectors are designed to

eliminate the period and carryover effects, respectively, and are given by:

$$\frac{1}{2}(-1 \ 0 \ 1 \ 0 \ 1 \ 0 \ -1 \ 0), \quad \frac{1}{2}(-2 \ 1 \ 0 \ 1 \ 2 \ -1 \ 0 \ -1)$$

These contrasts simultaneously eliminate both the period and carryover effects, resulting in the constraint:

$$\frac{1}{2} \begin{pmatrix} -1 & 0 & 1 & 0 & 1 & 0 & -1 & 0 \\ -2 & 1 & 0 & 1 & 2 & -1 & 0 & -1 \end{pmatrix} \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Through this formulation, the effects of period and carryover are constrained to be zero, thereby reducing the model's complexity. To evaluate how well this restricted model explains the data relative to the saturated model M_2 , a Bayes factor is computed. The Bayes factor comparing M_0 to M_2 is denoted by B_{02} and is expressed as:

$$B_{02} = \frac{3}{2q} \left(1 + \frac{3\hat{\theta}^2 - 2\hat{\theta}\hat{\lambda} + \hat{\lambda}^2}{qSS_2} \right)^{-3N/2} \quad (4.60)$$

In this expression, SS_2 denotes the within-subject sum of squares, and $\hat{\theta}$ and $\hat{\lambda}$ are the least squares estimates of the period and carryover effects, respectively. The quadratic form in the numerator quantifies the combined deviation from the null hypotheses $\theta = 0$ and $\lambda = 0$. The multiplicative constant $\frac{3}{2q}$ arises from the calibration procedure based on the imaginary training sample, ensuring that undefined prior constants are removed.

Having derived closed-form expressions for the Bayes factors comparing each of the reduced models M_{11} , M_{12} , and M_0 against the saturated model M_2 , it becomes possible to compare these reduced models directly. Rather than evaluating each pairwise comparison from scratch, a more elegant and computationally efficient route is available through the transitive property of Bayes factors.

This property follows directly from the definition of the Bayes factor between any two models

M_i and M_j , expressed as:

$$B_{ij} = \frac{\mathbb{P}(M_i | Y) \mathbb{P}(M_j)}{\mathbb{P}(M_j | Y) \mathbb{P}(M_i)}$$

and likewise, the Bayes factor comparing M_j and M_k is

$$B_{jk} = \frac{\mathbb{P}(M_j | Y) \mathbb{P}(M_k)}{\mathbb{P}(M_k | Y) \mathbb{P}(M_j)}$$

Multiplying the two expressions gives

$$B_{ij}B_{jk} = \frac{\mathbb{P}(M_i | Y) \mathbb{P}(M_j) \mathbb{P}(M_j | Y) \mathbb{P}(M_k)}{\mathbb{P}(M_j | Y) \mathbb{P}(M_i) \mathbb{P}(M_k | Y) \mathbb{P}(M_j)}$$

which simplifies by cancellation to

$$B_{ij}B_{jk} = \frac{\mathbb{P}(M_i | Y) \mathbb{P}(M_k)}{\mathbb{P}(M_k | Y) \mathbb{P}(M_i)} = B_{ik}$$

This derivation formally demonstrates that Bayes factors are transitive: the factor comparing M_i to M_k can be obtained by multiplying the Bayes factor comparing M_i to M_j with that comparing M_j to M_k . This identity holds regardless of prior model probabilities; however, under the commonly assumed condition of equal priors, the prior terms cancel and the expression reduces to a pure ratio of posterior probabilities.

This property, along with the reciprocal identity

$$B_{ij} = \frac{1}{B_{ji}}$$

enables indirect computation of Bayes factors between any pair of reduced models, provided each has been compared to a common model such as M_i . Using this framework, and the previously derived expressions for $B_{11,2}$, $B_{12,2}$ and $B_{0,2}$, direct Bayes factor comparisons between the reduced models can now be computed.

The Bayes factor comparing model M_{12} (with the period effect excluded) to model M_{12} (with

the carryover effect excluded) is:

$$B_{12,11} = \left(\frac{SS_2 + 2\hat{\lambda}^2/(3q)}{SS_2 + 2\hat{\theta}^2/q} \right)^{-3N/2} \quad (4.61)$$

This expression allows the strength of evidence in favor of one effect over the other to be assessed. Specifically, if this quantity exceeds one, it suggests that the data favor the presence of a carryover effect over a period effect, whereas a value less than one supports the opposite. The quantities $\hat{\lambda}^2$ and $\hat{\theta}^2$ are the squared estimates of the carryover and period effects, respectively, and the constants q and SS_2 denote the scaling term from the covariance structure and the within-subject sum of squares.

Next, the Bayes factor comparing the null model M_0 , which excludes both the period and carryover effects, to model M_{11} , which excludes only the period effect, is given by:

$$B_{0,11} = \left(\frac{3}{2q} \right)^{1/2} \left(\frac{SS_2 + 3\hat{\theta}^2/q - 2\hat{\theta}\hat{\lambda}/q + \hat{\lambda}^2/q}{SS_2 + 2\hat{\theta}^2/q} \right)^{-3N/2} \quad (4.62)$$

This form captures the evidence for removing the carryover effect from a model that has already excluded the period effect. The numerator includes a full quadratic combination of both $\hat{\theta}$ and $\hat{\lambda}$, which accounts for their joint variation and potential correlation. The square root prefactor arises from the prior normalizing constants when moving from one model dimension to another.

Likewise, the Bayes factor comparing the null model M_0 to model M_{12} , which excludes only the carryover effect, is given by:

$$B_{0,12} = \left(\frac{3}{2q} \right)^{1/2} \left(\frac{SS_2 + 3\hat{\theta}^2/q - 2\hat{\theta}\hat{\lambda}/q + \hat{\lambda}^2/q}{SS_2 + 2\hat{\lambda}^2/3q} \right)^{-3N/2} \quad (4.63)$$

This comparison examines the additional impact of removing the period effect from a model that already omits carryover. The numerator is identical to that in the previous expression, reflecting the shared structure of the null model, while the denominator now reflects the

specific contribution of the carryover term alone.

After deriving explicit expressions for the Bayes factors comparing each reduced model to the saturated model M_2 , the next natural step is to determine the posterior probability of each model given the observed data. These probabilities provide an interpretable, probability-based summary of which model is best supported, taking into account both the fit to data and any prior beliefs about the models.

Let us denote the four competing models as M_0 , M_{11} , M_{12} and M_2 , where M_2 is the saturated model that includes all effects. Each Bayes factor B_{i2} compares a model M_i to this reference model M_2 . To compute the posterior model probabilities, one must also consider the prior odds of each model relative to M_2 , which are denoted by:

$$\kappa_{i2} = \frac{\mathbb{P}(M_i)}{\mathbb{P}(M_2)}$$

According to the definition of a Bayes factor, this relationship holds:

$$B_{i2}\kappa_{i2} = \frac{\mathbb{P}(M_i|Y)}{\mathbb{P}(M_2|Y)}$$

Explicitly, for the three reduced models:

$$B_{02}\kappa_{02} = \frac{\mathbb{P}(M_0|Y)}{\mathbb{P}(M_2|Y)}$$

$$B_{11,2}\kappa_{11,2} = \frac{\mathbb{P}(M_{11}|Y)}{\mathbb{P}(M_2|Y)}$$

$$B_{12,2}\kappa_{12,2} = \frac{\mathbb{P}(M_{12}|Y)}{\mathbb{P}(M_2|Y)}$$

Since these posterior probabilities must sum to one across all four models, the posterior for

the saturated model M_2 can be written as:

$$\mathbb{P}(M_2|Y) = \frac{1}{1 + B_{0,2}\kappa_{02} + B_{11,2}\kappa_{11,2} + B_{12,2}\kappa_{12,2}}$$

Using this as a normalization constant, the remaining posterior probabilities follow directly:

$$\mathbb{P}(M_{11}|Y) = \frac{B_{11,2}\kappa_{11,2}}{1 + B_{02}\kappa_{02} + B_{11,2}\kappa_{11,2} + B_{12,2}\kappa_{12,2}}$$

$$\mathbb{P}(M_{12}|Y) = \frac{B_{12,2}\kappa_{12,2}}{1 + B_{02}\kappa_{02} + B_{11,2}\kappa_{11,2} + B_{12,2}\kappa_{12,2}}$$

$$\mathbb{P}(M_0|Y) = \frac{B_{0,2}\kappa_{02}}{1 + B_{02}\kappa_{02} + B_{11,2}\kappa_{11,2} + B_{12,2}\kappa_{12,2}}$$

This approach transforms the relative support provided by the Bayes factors into full posterior probabilities that sum to one, making it easier to compare all models simultaneously rather than pairwise. The use of the saturated model M_2 as the common reference ensures consistency across all comparisons.

To carry out these calculations, one must specify either the prior odds ratios κ_{ij} or the individual prior model probabilities $\mathbb{P}(M_i)$. If no strong prior belief favors any one model over the others, a common choice is to assign equal prior probabilities, so that:

$$\mathbb{P}(M_0) = \mathbb{P}(M_{11}) = \mathbb{P}(M_{12}) = \mathbb{P}(M_2) = 0.25$$

which implies:

$$\kappa_{02} = \kappa_{11,2} = \kappa_{12,2} = 1$$

This assumption reflects prior indifference, which means each model is considered equally plausible before observing the data. Under this assumption, the posterior expressions simplify:

$$\mathbb{P}(M_2 | Y) = \frac{1}{1 + B_{02} + B_{11,2} + B_{12,2}}$$

$$\mathbb{P}(M_2 | Y) = \frac{B_{02}}{1 + B_{02} + B_{11,2} + B_{12,2}}$$

$$\mathbb{P}(M_2 | Y) = \frac{B_{11,2}}{1 + B_{02} + B_{11,2} + B_{12,2}}$$

$$\mathbb{P}(M_2 | Y) = \frac{B_{12,2}}{1 + B_{02} + B_{11,2} + B_{12,2}}$$

If the analyst instead has reason to believe one model is more plausible apriori, these prior weights can be adjusted accordingly. In Grieve’s original framework, it was suggested that assuming indifference among models, using equal prior probabilities, is a reasonable starting point. In that case, the prior odds become one, and the model comparison is driven purely by the data through the Bayes factors.

By converting Bayes factors into posterior probabilities, a more intuitive and interpretable summary of model support is obtained. Rather than comparing models two at a time, this approach produces a complete probabilistic assessment over all models simultaneously. As a result, one can assess not only which model is most supported, but also how uncertain that preference is, based on how the posterior mass is distributed across the model space. This final step completes the Bayesian model selection framework for the two-period crossover design under consideration.

Bayesian model selection was applied to the ADAS-Cog dataset to compare four competing models: the saturated model M_2 , the null model M_0 , and two reduced models M_{11} and M_{12} , which exclude the period and carryover effects, respectively. To evaluate the relative support for each model, Bayes factors were calculated based on the observed data. These results are summarized in Table 4.6: While M_{12} showed the highest support among the reduced models, particularly when compared to M_{11} , the differences in Bayes factors across models were modest. According to commonly accepted interpretation guidelines, these values do not indicate strong or decisive evidence in favor of any single model.

Correspondingly, the posterior model probabilities under the assumption of equal prior weights are calculated and presented in Table 4.7: Each model was initially assigned equal

Table 4.6: Bayes factors comparing models

Bayes Factor	Value
$B_{11,2}$	1.027
$B_{12,2}$	3.602
B_{02}	2.708
$B_{12,11}$	3.507
$B_{0,11}$	2.637
$B_{0,12}$	0.752

Table 4.7: Posterior model probabilities for ADAS-Cog data

Model	Posterior Probability
$\mathbb{P}(M_2 Y)$	0.120
$\mathbb{P}(M_{11} Y)$	0.123
$\mathbb{P}(M_{12} Y)$	0.432
$\mathbb{P}(M_0 Y)$	0.325

prior probability, reflecting an unbiased stance regarding model structure. However, the resulting posterior model probabilities were relatively close, indicating no single model received overwhelming support. As such, it was not possible to confidently select one model over the others, and a model-averaging approach was adopted to incorporate this uncertainty into the inference for the treatment effect.

To this end, using Tables 4.4 and 4.7 and the total law of probability, the unconditional posterior probability of a positive treatment effect is computed by averaging over all four models, weighted by their respective posterior model probabilities:

$$\begin{aligned}
\mathbb{P}(\tau > 0 | Y) &= \mathbb{P}(\tau > 0 | Y, M_2)\mathbb{P}(M_2 | Y) \\
&\quad + \mathbb{P}(\tau > 0 | Y, M_{11})\mathbb{P}(M_{11} | Y) \\
&\quad + \mathbb{P}(\tau > 0 | Y, M_{12})\mathbb{P}(M_{12} | Y) \\
&\quad + \mathbb{P}(\tau > 0 | Y, M_0)\mathbb{P}(M_0 | Y) \\
&= 0.6117
\end{aligned}$$

This result suggests moderate support for a beneficial treatment effect, with the strength of evidence tempered by the presence of model uncertainty. Unlike traditional single-model approaches, this Bayesian strategy yields an inference that reflects both parameter-level and model-level uncertainty, offering a more robust and transparent conclusion.

4.3 Conclusion

Combining the full Bayesian formulation with a numerical application to ADAS-Cog scores from 41 participants, we examined the posterior distributions of the treatment effect τ under four competing models that differ in their assumptions about the carryover effect λ and the period/baseline offset θ . Closed-form posteriors were obtained under each model, and the resulting densities were plotted to assess how structural assumptions shape inference. The probabilities $\mathbb{P}(\tau > 0 \mid Y, M)$ varied substantially across models, from 0.47 under the saturated model M_2 to 0.79 under the reduced model M_{11} (which sets $\theta = 0$), illustrating how constraining nuisance parameters can amplify or attenuate apparent evidence for a treatment effect.

To account for uncertainty about the correct model, we compared models using Bayes factors and derived posterior model probabilities. Although M_{12} (which sets $\lambda = 0$, i.e., no carryover) achieved the largest posterior probability (0.432), the remaining models also retained non-negligible support, with none clearly dominating. Given equal prior model weights and this relatively diffuse posterior support, we adopted Bayesian model averaging rather than selecting a single structure. The resulting unconditional probability $\mathbb{P}(\tau > 0 \mid Y) = 0.6117$ indicates moderate evidence of a beneficial treatment effect while fully incorporating model uncertainty. Taken together, these results emphasize that conclusions about τ depend on assumptions about θ and λ , and they highlight the value of a Bayesian framework for robust, interpretable inference in crossover trials.

Chapter 5

Discussion and Conclusion

Summary. This chapter discusses the results of the study and presents the main conclusions drawn from the analyses. The findings from the classical, bootstrap, and Bayesian methods are interpreted together to provide a clearer understanding of the potential effects of transcranial alternating current stimulation (tACS) combined with cognitive exercise on cognitive performance in individuals with dementia. By considering these different analytical approaches, the chapter highlights how each method contributes to understanding the treatment effect and the uncertainty in the results.

The chapter also discusses the limitations of the study, including the relatively small sample size, the possible presence of carryover effects in the crossover design, and the limitations of using ADAS-Cog as a global measure of cognitive function. Finally, possible directions for future research are presented, emphasizing the importance of larger studies, more detailed cognitive assessments, and improved analytical approaches to further explore the potential role of neuromodulation interventions in dementia research.

5.1 Discussion

This study examined the effects of tACS combined with cognitive exercise within a randomized 2×2 crossover framework. Rather than aiming to establish definitive evidence of efficacy, the analytical strategy was designed to explore how different inferential perspectives characterize treatment effects and uncertainty in a setting where sample size, outcome variability, and model complexity impose inherent limitations on statistical inference. The discussion therefore focuses on interpreting the findings within these constraints rather than reiterating the numerical results presented earlier.

A central consideration in interpreting the present findings is the extent to which statistical inference depends on the assumptions underlying the analytical models. In crossover trials, conclusions about treatment effects are closely linked to assumptions regarding period effects, carryover effects, and potential interaction structures. When sample sizes are relatively small, as is often the case in dementia research, even modest deviations from these assumptions can influence the resulting estimates and the interpretation of treatment effects. The analyses conducted in this thesis demonstrate that uncertainty regarding model structure should not be viewed merely as a technical issue, but rather as an integral component of statistical inference that must be acknowledged when interpreting results.

The use of multiple analytical frameworks further illustrates how different statistical approaches emphasize distinct inferential goals. Classical methods primarily focus on hypothesis testing under fixed model assumptions and provide a familiar but often conservative framework for evaluating treatment effects. Resampling-based approaches, such as the bootstrap, shift attention toward the stability and variability of estimates under repeated sampling, offering insight into uncertainty that may not be fully captured by parametric models. Bayesian methods, in contrast, represent uncertainty explicitly through probability distributions and enable direct probabilistic statements about treatment effects while accounting for model uncertainty. When considered together, these approaches provide complementary per-

spectives that contribute to a more comprehensive understanding of the evidence contained in the data.

Importantly, the presence of uncertainty across analytical frameworks should not be interpreted as methodological inconsistency or analytical weakness. Rather, it reflects the inherent difficulty of detecting small or heterogeneous treatment effects in cognitively impaired populations over relatively short intervention periods. Cognitive outcomes such as those evaluated in this study are influenced by numerous biological, behavioral, and contextual factors, many of which cannot be fully controlled or observed within clinical trials. In this context, the absence of clear or consistent statistical evidence is itself informative and highlights the need for cautious interpretation of treatment effects.

From a methodological perspective, this work demonstrates the value of explicitly acknowledging uncertainty rather than attempting to minimize it through reliance on a single inferential framework. By examining the data through classical, resampling-based, and Bayesian approaches, the analysis provides a transparent representation of the range of conclusions supported by the observed data. Such an approach is particularly useful in exploratory or early-stage clinical studies, where the objective is often to inform future research rather than to establish definitive evidence of treatment efficacy.

From a clinical standpoint, the findings also illustrate the challenges associated with evaluating non-pharmacological interventions for dementia within constrained study designs. While crossover trials provide practical and ethical advantages by allowing each participant to serve as their own control, short-term designs may have limited ability to capture delayed, cumulative, or subgroup-specific effects. Consequently, modest or uncertain findings should be interpreted as indications that further investigation is warranted rather than as definitive evidence regarding the clinical effectiveness of the intervention.

Overall, these considerations highlight the importance of aligning analytical strategies with the practical realities of clinical research in vulnerable populations. In small crossover

trials, uncertainty should be treated not as an obstacle to interpretation but as an essential feature that must be explicitly modeled and communicated when evaluating treatment effects.

The following section summarizes the key findings of the study and synthesizes the conclusions that can be drawn from the classical, bootstrap, and Bayesian analyses presented in this thesis.

5.2 Conclusions

This thesis investigated the potential cognitive effects of tACS combined with cognitive exercise in individuals with dementia using a randomized two-period, two-treatment (2×2) crossover design. Participants received both real and sham stimulation across two study periods, allowing each participant to serve as their own control and thereby reducing between-subject variability. Cognitive outcomes were evaluated using the Alzheimer’s Disease Assessment Scale–Cognitive Subscale (ADAS-Cog), a widely used measure of global cognitive performance in dementia research.

To evaluate the intervention comprehensively, three complementary analytical frameworks were applied. Classical statistical approaches, including ANOVA- and ANCOVA-based models as well as non-parametric procedures, were first used to estimate treatment effects and test hypotheses under standard modeling assumptions. A nonparametric bootstrap analysis was then conducted to assess the stability of the treatment estimates under repeated resampling and to reduce reliance on strict parametric assumptions given the modest sample size. Finally, Bayesian model comparison techniques were implemented to provide probabilistic statements about treatment benefit and to evaluate the relative support for competing model structures that incorporate treatment, period, and carryover effects.

Across these analytical frameworks, the findings present a mixed but informative

picture of the potential effect of tACS on ADAS-Cog performance. Classical analyses did not provide statistically significant evidence of a difference between real and sham stimulation. Even after adjusting for baseline cognitive scores to reduce between-subject variability, the estimated treatment effect remained small and statistically non-significant. Non-parametric tests produced similar conclusions, indicating no clear improvement associated with real tACS.

The bootstrap analysis supported these results by showing that the treatment estimate remained centered near the observed statistic across repeated resampled datasets. The resulting bootstrap inference suggested that the data did not strongly contradict the null hypothesis of no treatment effect, indicating that the lack of statistical significance observed in classical analyses was unlikely to be solely attributable to model assumptions.

In contrast, the Bayesian analysis provided a slightly different perspective by allowing uncertainty to be expressed probabilistically and by formally accounting for model uncertainty. When competing model structures were compared and averaged, the unconditional posterior probability of a beneficial treatment effect was estimated as $\mathbb{P}(\tau > 0 \mid Y) = 0.6117$. While this probability does not constitute strong evidence of efficacy, it suggests moderate support for a potential improvement associated with real tACS that was less apparent in the classical and bootstrap analyses.

Taken together, these results suggest that any potential cognitive benefit of tACS on global ADAS-Cog performance is likely modest and not consistently detectable across all inferential approaches. Classical and bootstrap analyses point toward negligible treatment differences, whereas the Bayesian framework indicates moderate but inconclusive evidence in favor of improvement. These mixed findings highlight both the complexity of analyzing small crossover trials and the challenges associated with detecting subtle cognitive effects in dementia populations.

5.3 Limitations

Several limitations should be considered when interpreting the findings of this study. First, although the crossover design provides strong control over between-subject variability and improves statistical efficiency, the overall sample size remained modest. With 41 participants completing assessments across two periods, the study may have been underpowered to detect small or heterogeneous treatment effects, which are common in dementia interventions. This limited statistical power likely contributed to the mixed results observed across the classical and bootstrap analyses.

Second, while the washout interval between treatment periods ranged from several weeks to as long as two months, which is generally considered sufficient to reduce residual stimulation effects, the possibility of subtle or individual-specific carryover effects cannot be entirely excluded. Even minimal residual influences, if present, could reduce the observable differences between real and sham periods and contribute to increased variability in treatment estimates.

Third, although the ADAS-Cog is widely used and clinically meaningful, it represents a broad global measure of cognitive functioning and may not fully capture short-term or domain-specific changes associated with tACS combined with cognitive training. It is therefore possible that the intervention affected specific cognitive processes that were not fully reflected in the overall ADAS-Cog score. Future studies incorporating additional domain-specific cognitive measures may provide a more comprehensive assessment of potential treatment effects.

Finally, the present analysis did not explicitly account for factors such as daily fluctuations in cognitive performance, fatigue, attention, or participant engagement during cognitive training sessions. These factors can introduce additional variability in dementia populations and may mask subtle treatment-related changes that could otherwise be detectable under

more controlled conditions.

5.4 Future Research Directions

Future research should consider incorporating additional sources of information that may influence cognitive performance but were not explicitly modeled in the present study. Factors such as fatigue, sleep quality, daily functioning, medication timing, and participant engagement may introduce meaningful variability in dementia populations. Collecting and integrating such measures into future analyses could improve the interpretation of cognitive outcomes and help distinguish treatment-related changes from normal day-to-day fluctuations.

Another important direction involves the development of adaptive or individualized stimulation strategies. Differences in neural characteristics and cognitive profiles among participants may influence responsiveness to tACS. Approaches that incorporate baseline EEG characteristics or individualized cognitive assessments could help tailor stimulation parameters, such as frequency or phase, to the neural dynamics of each participant. The use of machine learning or data-driven modeling techniques may further support the development of personalized stimulation protocols and provide insight into why certain individuals respond more strongly to neuromodulation interventions.

From an analytical perspective, future work may benefit from applying more flexible statistical frameworks that better accommodate the complexity of crossover designs and small clinical samples. Hierarchical Bayesian models, individualized treatment-response modeling, and joint modeling approaches that combine behavioral and neurophysiological data represent promising avenues. These approaches may allow for improved estimation of treatment effects, more effective modeling of inter-individual variability, and deeper understanding of the mechanisms underlying potential cognitive changes associated with neuromodulation

interventions.

5.5 Final Remarks

This thesis investigated the potential cognitive effects of tACS combined with cognitive exercise using a randomized two-period crossover design. The analysis incorporated classical statistical methods, bootstrap resampling, and Bayesian modeling in order to examine the robustness of the findings across multiple inferential frameworks. While the classical and resampling approaches did not demonstrate a clear improvement in ADAS-Cog performance, the Bayesian analysis provided moderate probabilistic support for a beneficial treatment effect.

These findings reflect both the inherent variability present in dementia research and the methodological challenges associated with detecting subtle cognitive changes in relatively small clinical samples. Although the results do not provide definitive evidence of treatment efficacy, they highlight the importance of using complementary analytical approaches when evaluating complex clinical data.

Beyond the specific findings related to tACS, this work contributes methodologically by demonstrating how classical, resampling-based, and Bayesian perspectives can be integrated to provide a more comprehensive interpretation of treatment effects in crossover trials. Such an approach may be particularly valuable in early-stage clinical research, where uncertainty is unavoidable and careful interpretation of evidence is essential.

Overall, the study provides insight into the potential role of neuromodulation interventions in dementia research while emphasizing the need for continued investigation using larger samples, refined outcome measures, and advanced analytical strategies.

Bibliography

- [1] Byron Jones and Michael G Kenward. *Design and Analysis of Cross-Over Trials*. CRC Press, 3rd edition, 2014. doi: 10.1201/B17537/DESIGN-ANALYSIS-CROSS-TRIALS-MICHAEL-KENWARD-BYRON-JONES/RIGHTS-AND-PERMISSIONS.
- [2] Stephen Senn. *Crossover Trials in Clinical Research*. Wiley, Chichester, UK, 2 edition, 2002.
- [3] Philip Sedgwick. What is a crossover trial? *BMJ*, 348:g3191, 2014. doi: 10.1136/bmj.g3191. URL <https://doi.org/10.1136/bmj.g3191>.
- [4] John F Reed. Analysis of two-treatment, two-period crossover trials in emergency medicine. *Annals of Emergency Medicine*, 43(1):54–58, 2004. doi: 10.1016/S0196-0644(03)00661-9.
- [5] Norman D Smith. Discussion of ‘estimating treatment effects in clinical crossover trials’: A regulatory perspective. *Journal of Biopharmaceutical Statistics*, 8(2):243–247, 1998. doi: 10.1080/10543409808835236.
- [6] James E Grizzle. The two-period change-over design and its use in clinical trials. *Biometrics*, 21(2):467, 1965. doi: 10.2307/2528104.
- [7] Peter R Freeman. The performance of the two-stage analysis of two-treatment, two-

- period crossover trials. *Statistics in Medicine*, 8(12):1421–1432, 1989. doi: 10.1002/SIM.4780081202.
- [8] Richard P Kershner and Walter T Federer. Two-treatment crossover designs for estimating a variety of effects. *Journal of the American Statistical Association*, 76(375):612–619, 1981. doi: 10.1080/01621459.1981.10477693.
- [9] Joseph L Fleiss. A critique of recent research on the two-treatment crossover design. *Controlled Clinical Trials*, 10(3):237–243, 1989. doi: 10.1016/0197-2456(89)90065-2.
- [10] Alzheimer’s Association. 2015 alzheimer’s disease facts and figures. *Alzheimer’s & Dementia*, 11(3):332–384, 2015. doi: 10.1016/j.jalz.2015.02.003.
- [11] C. N. Harada, M. C. Natelson Love, and K. L. Triebel. Normal cognitive aging. *Clinics in Geriatric Medicine*, 29(4):737–752, 2013. doi: 10.1016/j.cger.2013.07.002.
- [12] G. A. Marshall, D. M. Rentz, M. T. Frey, J. J. Locascio, K. A. Johnson, R. A. Sperling, and Alzheimer’s Disease Neuroimaging Initiative. Executive function and instrumental activities of daily living in mild cognitive impairment and alzheimer’s disease. *Alzheimer’s & Dementia: The Journal of the Alzheimer’s Association*, 7(3):300–308, 2011. doi: 10.1016/j.jalz.2010.04.005.
- [13] T. Berman and A. Bayati. What are neurodegenerative diseases and how do they affect the brain? *Frontiers for Young Minds*, 6:70, 2018. doi: 10.3389/frym.2018.00070.
- [14] S. Karantzoulis and J. E. Galvin. Distinguishing alzheimer’s disease from other major forms of dementia. *Expert Review of Neurotherapeutics*, 11(11):1579–1591, 2011. doi: 10.1586/ern.11.155.
- [15] S. G. Di Santo, F. Prinelli, F. Adorni, C. Caltagirone, and M. Musicco. A meta-analysis of the efficacy of donepezil, rivastigmine, galantamine, and memantine in relation to

- severity of alzheimer’s disease. *Journal of Alzheimer’s Disease*, 35(2):349–361, 2013. doi: 10.3233/JAD-122140.
- [16] M. Zvěřová. Clinical aspects of alzheimer’s disease. *Clinical Biochemistry*, 72:3–6, 2019. doi: 10.1016/j.clinbiochem.2019.04.015.
- [17] K. G. Yiannopoulou and S. G. Papageorgiou. Current and future treatments in alzheimer disease: An update. *Journal of Central Nervous System Disease*, 12: 1179573520907397, 2020. doi: 10.1177/1179573520907397.
- [18] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S. G. Costafreda, A. Dias, N. Fox, L. N. Gitlin, R. Howard, H. C. Kales, M. Kivimäki, E. B. Larson, A. Ogunniyi, V. Orgeta, and N. Mukadam. Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet*, 396(10248):413–446, 2020. doi: 10.1016/S0140-6736(20)30367-6.
- [19] S. Salloway, R. Sperling, N. C. Fox, K. Blennow, W. Klunk, M. Raskind, M. Sabbagh, L. S. Honig, A. P. Porsteinsson, S. Ferris, M. Reichert, N. Ketter, B. Nejadnik, V. Guenzler, M. Miloslavsky, D. Wang, Y. Lu, J. Lull, I. C. Tudor, E. Liu, Bapineuzumab 301, and 302 Clinical Trial Investigators. Two phase 3 trials of bapineuzumab in mild-to-moderate alzheimer’s disease. *The New England Journal of Medicine*, 370(4):322–333, 2014. doi: 10.1056/NEJMoa1304839.
- [20] K. L. Lanctôt, N. Herrmann, K. K. Yau, L. R. Khan, B. A. Liu, M. M. LouLou, and T. R. Einarson. Efficacy and safety of cholinesterase inhibitors in alzheimer’s disease: A meta-analysis. *CMAJ: Canadian Medical Association Journal*, 169(6):557–564, 2003.
- [21] Z. Moussavi, K. Kimura, L. Kehler, C. de Oliveira Francisco, and B. Lithgow. A novel program to improve cognitive function in individuals with dementia using transcranial

- alternating current stimulation (tacs) and tutored cognitive exercises. *Frontiers in Aging*, 2:632545, 2021. doi: 10.3389/fragi.2021.632545.
- [22] I. Gonsalvez, R. Baror, P. Fried, E. Santarnecchi, and A. Pascual-Leone. Therapeutic noninvasive brain stimulation in alzheimer’s disease. *Current Alzheimer Research*, 14(4):362–376, 2017. doi: 10.2174/1567205013666160930113907.
- [23] Davide Maria Cammisuli, Augusto Innocenti, Ferdinando Franzoni, and Carlo Pruneti. Aerobic exercise effects upon cognition in mild cognitive impairment: A systematic review of randomized controlled trials. *Archives Italiennes de Biologie*, 155(1–2):54–62, 7 2017. doi: 10.12871/000398292017126. URL <https://pubmed.ncbi.nlm.nih.gov/28715598/>.
- [24] Ewelina Maliszewska-Cyna, Madelaine Lynch, Jonathan Oore, Paul Nagy, and Isabelle Aubert. The benefits of exercise and metabolic interventions for the prevention and early treatment of alzheimer’s disease. *Current Alzheimer Research*, 14(1):47–60, 8 2017. doi: 10.2174/1567205013666160819125400. URL <https://pubmed.ncbi.nlm.nih.gov/27539597/>.
- [25] Bushra Imtiaz, Anna Maija Tolppanen, Miia Kivipelto, and Hilka Soininen. Future directions in alzheimer’s disease from risk factors to prevention. *Biochemical Pharmacology*, 88(4):661–670, 4 2014. doi: 10.1016/j.bcp.2014.01.003. URL <https://pubmed.ncbi.nlm.nih.gov/24418410/>.
- [26] Zahra Moussavi, Grant Rutherford, Brian Lithgow, Colleen Millikin, Mandana Modirrousta, Behzad Mansouri, Xikui Wang, Craig Omelan, Lesley Fellows, Paul Fitzgerald, and Lisa Koski. Repeated transcranial magnetic stimulation for improving cognition in patients with alzheimer disease: Protocol for a randomized, double-blind, placebo-controlled trial. *JMIR Research Protocols*, 10(1), 1 2021. doi: 10.2196/25144. URL <https://pubmed.ncbi.nlm.nih.gov/33416500/>.

- [27] Zahra Jafari, Bryan E. Kolb, and Majid H. Mohajerani. Neural oscillations and brain stimulation in alzheimer’s disease. *Progress in Neurobiology*, 194:101878, 11 2020. doi: 10.1016/j.pneurobio.2020.101878. URL <https://pubmed.ncbi.nlm.nih.gov/32615147/>.
- [28] A. J. Woods, A. Antal, M. Bikson, P. S. Boggio, A. R. Brunoni, P. Celnik, L. G. Cohen, F. Fregni, C. S. Herrmann, E. S. Kappenman, H. Knotkova, D. Liebetanz, C. Miniussi, P. C. Miranda, W. Paulus, A. Priori, D. Reato, C. Stagg, N. Wenderoth, and M. A. Nitsche. A technical guide to tdc, and related non-invasive brain stimulation tools. *Clinical Neurophysiology*, 127(2):1031–1048, 2 2016. doi: 10.1016/j.clinph.2015.11.012. URL <https://pubmed.ncbi.nlm.nih.gov/26652115/>.
- [29] Christoph S. Herrmann, Stefan Rach, Toralf Neuling, and Daniel Strüber. Transcranial alternating current stimulation: A review of the underlying mechanisms and modulation of cognitive processes. *Frontiers in Human Neuroscience*, 7(5):279, 5 2013. doi: 10.3389/fnhum.2013.00279. URL <https://pubmed.ncbi.nlm.nih.gov/23785325/>.
- [30] Annamaria Antal and Walter Paulus. Transcranial alternating current stimulation (tacs). *Frontiers in Human Neuroscience*, 7(JUN):317, 6 2013. doi: 10.3389/fnhum.2013.00317. URL <https://www.frontiersin.org/articles/10.3389/fnhum.2013.00317>.
- [31] Daria Antonenko, Miriam Fixel, Ulrike Grittner, Michal Lavidor, and Agnes Flöel. Effects of transcranial alternating current stimulation on cognitive functions in healthy young and older adults. *Neural Plasticity*, 2016:4274127, 2016. doi: 10.1155/2016/4274127.
- [32] J. Kim, H. Kim, H. Jeong, D. Roh, and D. H. Kim. tacs as a promising therapeutic option for improving cognitive function in mild cognitive impairment: A direct comparison between tacs and tdc. *Journal of Psychiatric Research*, 141:248–256,

- 9 2021. doi: 10.1016/j.jpsychires.2021.07.012. URL <https://doi.org/10.1016/j.jpsychires.2021.07.012>.
- [33] Omar Elyamany, Gregor Leicht, Christoph S. Herrmann, and Christoph Mulert. Transcranial alternating current stimulation (tacs): From basic mechanisms towards first applications in psychiatry. *European Archives of Psychiatry and Clinical Neuroscience*, 271(1):135–156, 2021. doi: 10.1007/s00406-020-01209-9. URL <https://doi.org/10.1007/s00406-020-01209-9>.
- [34] Samantha J. Booth, Jason R. Taylor, Laura J. E. Brown, and Gorana Pobric. The effects of transcranial alternating current stimulation on memory performance in healthy adults: A systematic review. *Cortex*, 147:112–139, 2 2022. doi: 10.1016/j.cortex.2021.12.001. URL <https://pubmed.ncbi.nlm.nih.gov/35032750/>.
- [35] Katharina Klink, Sven Paßmann, Florian H. Kasten, and Jessica Peter. The modulation of cognitive performance with transcranial alternating current stimulation: A systematic review of frequency-specific effects. *Brain Sciences*, 10(12):932, 2020. doi: 10.3390/brainsci10120932. URL <https://pubmed.ncbi.nlm.nih.gov/33276533/>.
- [36] M. Dhaynaut, D. Cappon, Rachel Paciorek, Joanna Macone, A. Connor, N. Guehl, Á. Pascual-Leone, G. Fakhri, and E. Santarnecchi. Effects of modulating gamma oscillations via 40hz transcranial alternating current stimulation (tacs) on tau pet imaging in mild to moderate alzheimer’s disease. *The Journal of Nuclear Medicine*, 2020. URL https://jnm.snmjournals.org/content/61/supplement_1/625. Abstract or conference presentation.
- [37] Kun Hu, Ruihan Wan, Ying Liu, Maolin Niu, Jianrui Guo, and Feng Guo. Effects of transcranial alternating current stimulation on motor performance and motor learning for healthy individuals: A systematic review and meta-analysis. *Front-*

- tiers in Physiology*, 13:1064584, 11 2022. doi: 10.3389/fphys.2022.1064584. URL <https://pubmed.ncbi.nlm.nih.gov/36467691/>.
- [38] Ao Guan, Shaoshuang Wang, Ailing Huang, Chenyue Qiu, Yansong Li, Xuying Li, Jinfei Wang, Qiang Wang, and Bin Deng. The role of gamma oscillations in central nervous system diseases: Mechanism and treatment. *Frontiers in Cellular Neuroscience*, 16:962957, 7 2022. doi: 10.3389/fncel.2022.962957. URL <https://pubmed.ncbi.nlm.nih.gov/35966207/>.
- [39] Arianna Menardi, Simone Rossi, Giacomo Koch, Harald Hampel, Andrea Vergallo, Michael A. Nitsche, Yaakov Stern, Barbara Borroni, Stefano F. Cappa, Maria Cotelli, Giulio Ruffini, Georges El-Fakhri, Paolo M. Rossini, Brad Dickerson, Andrea Antal, Claudio Babiloni, Jean Pascal Lefaucheur, Bruno Dubois, Gustavo Deco, Ulf Ziemann, Alvaro Pascual-Leone, and Emiliano Santarnecchi. Toward noninvasive brain stimulation 2.0 in alzheimer’s disease. *Ageing Research Reviews*, 75:101555, 3 2022. doi: 10.1016/j.arr.2021.101555. URL <https://pubmed.ncbi.nlm.nih.gov/34973457/>.
- [40] Artemis Traikapi and Nikos Konstantinou. Gamma oscillations in alzheimer’s disease and their potential therapeutic role. *Frontiers in Systems Neuroscience*, 15:782399, 12 2021. doi: 10.3389/fnsys.2021.782399. URL <https://pubmed.ncbi.nlm.nih.gov/34966263/>.
- [41] Alexandra S. Klein, José R. Donoso, Richard Kempter, Dietmar Schmitz, and Prateep Beed. Early cortical changes in gamma oscillations in alzheimer’s disease. *Frontiers in Systems Neuroscience*, 10:83, 10 2016. doi: 10.3389/fnsys.2016.00083. URL <https://pubmed.ncbi.nlm.nih.gov/27833535/>.
- [42] Laure Verret, Edward O. Mann, Giao B. Hang, Albert M. I. Barth, Inma Cobos, Kaitlyn Ho, Nino Devidze, Eliezer Masliah, Anatol C. Kreitzer, Istvan Mody, Lennart Mucke, and Jorge J. Palop. Inhibitory interneuron deficit links altered network activity

- and cognitive dysfunction in alzheimer model. *Cell*, 149(3):708–721, 4 2012. doi: 10.1016/j.cell.2012.02.046. URL <https://pubmed.ncbi.nlm.nih.gov/22541439/>.
- [43] Romain Goutagny, Ning Gu, Chelsea Cavanagh, Jesse Jackson, Jean Guy Chabot, Rémi Quirion, Slavica Krantic, and Sylvain Williams. Alterations in hippocampal network oscillations and theta-gamma coupling arise before $a\beta$ overproduction in a mouse model of alzheimer’s disease. *The European Journal of Neuroscience*, 37(12): 1896–1902, 6 2013. doi: 10.1111/ejn.12233. URL <https://pubmed.ncbi.nlm.nih.gov/23773058/>.
- [44] Alexandra J. Mably, Brian J. Gereke, Dylan T. Jones, and Laura Lee Colgin. Impairments in spatial representations and rhythmic coordination of place cells in the 3xtg mouse model of alzheimer’s disease. *Hippocampus*, 27(4):378–392, 4 2017. doi: 10.1002/hipo.22697. URL <https://pubmed.ncbi.nlm.nih.gov/28032686/>.
- [45] Hannah F. Iaccarino, Annabelle C. Singer, Anthony J. Martorell, Andrii Rudenko, Fan Gao, Tyler Z. Gillingham, Hansruedi Mathys, Jinsoo Seo, Oleg Kritskiy, Fatema Abdurrob, Chinnakkaruppan Adaikkan, Rebecca G. Canter, Richard Rueda, Emery N. Brown, Edward S. Boyden, and Li-Huei Tsai. Gamma frequency entrainment attenuates amyloid load and modifies microglia. *Nature*, 540(7632):230–235, 12 2016. doi: 10.1038/nature20587. URL <https://pubmed.ncbi.nlm.nih.gov/27929004/>.
- [46] Anthony J. Martorell, Abigail L. Paulson, Ho Jun Suk, Fatema Abdurrob, Gabrielle T. Drummond, Webster Guan, Jennie Z. Young, David Nam Woo Kim, Oleg Kritskiy, Scarlett J. Barker, Vamsi Mangena, Stephanie M. Prince, Emery N. Brown, Kwanghun Chung, Edward S. Boyden, Annabelle C. Singer, and Li-Huei Tsai. Multi-sensory gamma stimulation ameliorates alzheimer’s-associated pathology and improves cognition. *Cell*, 177(2):256–271.e22, 4 2019. doi: 10.1016/j.cell.2019.02.014. URL <https://pubmed.ncbi.nlm.nih.gov/30879788/>.

- [47] Tomonori Nomura, Akihiko Asao, and Ayumi Kumasaka. Transcranial alternating current stimulation over the prefrontal cortex enhances episodic memory recognition. *Experimental Brain Research*, 237(7):1709–1715, 7 2019. doi: 10.1007/s00221-019-05543-w. URL <https://pubmed.ncbi.nlm.nih.gov/31011766/>.
- [48] A. Bahar-Fuchs, A. Martyr, A. M. Goh, J. Sabates, and L. Clare. Cognitive training for people with mild to moderate dementia. *The Cochrane Database of Systematic Reviews*, 2019(3):CD013069, 2019. doi: 10.1002/14651858.CD013069.pub2.
- [49] Alden L Gross, Jeanine M Parisi, Adam P Spira, Allison M Kueider, Jean Y Ko, Jane S Saczynski, Quincy M Samus, and George W Rebok. Memory training interventions for older adults: a meta-analysis. *Aging & Mental Health*, 16(6):722–734, 2012. doi: 10.1080/13607863.2012.667783.
- [50] S. L. Willis, S. L. Tennstedt, M. Marsiske, K. Ball, J. Elias, K. M. Koepke, J. N. Morris, G. W. Rebok, F. W. Unverzagt, A. M. Stoddard, and E. Wright. Long-term effects of cognitive training on everyday functional outcomes in older adults. *JAMA*, 296(23):2805–2814, 2006. doi: 10.1001/jama.296.23.2805. URL <https://doi.org/10.1001/jama.296.23.2805>.
- [51] G. E. Smith, P. Housen, K. Yaffe, R. Ruff, R. F. Kennison, H. W. Mahncke, and E. M. Zelinski. A cognitive training program based on principles of brain plasticity: Results from the improvement in memory with plasticity-based adaptive cognitive training (impact) study. *Journal of the American Geriatrics Society*, 57:594–603, 2009. doi: 10.1111/j.1532-5415.2008.02167.x. URL <https://doi.org/10.1111/j.1532-5415.2008.02167.x>.
- [52] J. D. Edwards, H. Xu, D. O. Clark, L. T. Guey, L. A. Ross, and F. W. Unverzagt. Speed of processing training results in lower risk of dementia. *Alzheimer’s & Dementia*:

- Translational Research & Clinical Interventions*, 3:603–611, 2017. doi: 10.1016/j.trci.2017.09.002. URL <https://doi.org/10.1016/j.trci.2017.09.002>.
- [53] B. S. A. Polat and A. C. Sarilar. The effects of computer-based cognitive training program on reaction times of patients with early stage alzheimer’s disease and traumatic brain injury. *Medical Science and Discovery*, 9(2):107–111, 2022. doi: 10.36472/MSD.V9I2.683. URL <https://doi.org/10.36472/MSD.V9I2.683>.
- [54] Nicola Gates, Perminder Sachdev, Maria Fiatarone Singh, and Michael Valenzuela. Cognitive and memory training in adults at risk of dementia: A systematic review. *BMC Geriatrics*, 11:55, 2011. doi: 10.1186/1471-2318-11-55.
- [55] Yoshio Kaneko and Matcheri Keshavan. Cognitive remediation in schizophrenia. *Clinical Psychopharmacology and Neuroscience*, 10(3):125–135, 2012. doi: 10.9758/cpn.2012.10.3.125.
- [56] Brita Fritsch, Janine Reis, Keri Martinowich, Heidi M Schambra, Yuan Yuan Ji, Leonardo G Cohen, and Bai Lu. Direct current stimulation promotes bdnf-dependent synaptic plasticity: Potential implications for motor learning. *Neuron*, 66(2):198–204, 2010. ISSN 0896-6273. doi: 10.1016/j.neuron.2010.03.035. URL <https://www.sciencedirect.com/science/article/pii/S0896627310002382>.
- [57] Yixuan Lv, Sheng Wu, Michael A Nitsche, Tao Yue, Volker R Zschorlich, and Fang Qi. A meta-analysis of the effects of transcranial direct current stimulation combined with cognitive training on working memory in healthy older adults. *Frontiers in Aging Neuroscience*, 16:1454755, 2024. doi: 10.3389/fnagi.2024.1454755.
- [58] Amir Hossein Zarifkar, Aida Zarifkar, Mohammad Nami, Amir Rafati, Hajar Aligholi, and Fatemeh Vafae. Ameliorative effects of different transcranial electrical stimulation paradigms on the novel object recognition task in a rat model of alzheimer disease. *Galen Medical Journal*, 8:e1440, 2019. doi: 10.31661/gmj.v8i0.1440.

- [59] Alberto Benussi, Valentina Cantoni, Maria Sofia Cotelli, Maria Cotelli, Chiara Brattini, Abhishek Datta, Chris Thomas, Emiliano Santarnecchi, Alvaro Pascual-Leone, and Barbara Borroni. Exposure to gamma tacs in alzheimer’s disease: A randomized, double-blind, sham-controlled, crossover, pilot study. *Brain Stimulation*, 14(3):531–540, 2021. ISSN 1935-861X. doi: 10.1016/j.brs.2021.03.007. URL <https://www.sciencedirect.com/science/article/pii/S1935861X21000589>.
- [60] Nathan Jacobson, Brian Lithgow, Mohammad Jafari Jozani, and Zahra Moussavi. The effect of transcranial alternating current stimulation with cognitive training on executive brain function in individuals with dementia: Protocol for a crossover randomized controlled trial. *JMIR Research Protocols*, 11(4):e37282, 2022. doi: 10.2196/37282.
- [61] Maria Anabel Uehara, Sumeet Kalia, Mari Garcia Campuzano, Mohammad Jafari Jozani, Brian Lithgow, and Zahra Moussavi. Cognitive and neuropsychiatric effects of 40 hz tacs simultaneously with cognitive exercises for dementia: A randomized, crossover, double-blind, sham-controlled study. *Medicina*, 61(4), 2025.
- [62] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005. doi: 10.1111/j.1532-5415.2005.53221.x.
- [63] Soterix Medical Inc. 1x1 transcranial alternating current stimulation (tacs) system, 2025. URL <https://soterixmedical.com/research/1x1/tacs>. Image retrieved from the Soterix Medical website.
- [64] Zahra Moussavi. Mindtriggers. <https://apps.apple.com/us/app/mind-triggers/id1508008769>, 2020. Mobile application. Accessed on January 20, 2025.
- [65] William G Rosen, Richard C Mohs, and Kenneth L Davis. A new rating scale for

- alzheimer's disease. *The American Journal of Psychiatry*, 141(11):1356–1364, 1984. doi: 10.1176/ajp.141.11.1356.
- [66] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2 edition, 2002. ISBN 9780534243128.
- [67] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, 2004. ISBN 9780387402727. doi: 10.1007/978-0-387-21736-9. URL <https://doi.org/10.1007/978-0-387-21736-9>.
- [68] John A. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/Cole, Belmont, CA, 3 edition, 2007. ISBN 9780495118670.
- [69] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968. doi: 10.1093/biomet/55.1.1. URL <https://doi.org/10.1093/biomet/55.1.1>.
- [70] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of shapiro–wilk, kolmogorov–smirnov, lilliefors, and anderson–darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.
- [71] Howard Levene. Robust tests for equality of variances. In Ingram Olkin, editor, *Contributions to Probability and Statistics*, pages 278–292. Stanford University Press, Stanford, CA, 1960.
- [72] Douglas C. Montgomery. *Design and Analysis of Experiments*. Wiley, Hoboken, NJ, 8 edition, 2013. ISBN 9781118146927.
- [73] Michael H. Kutner, Christopher J. Nachtsheim, and John Neter. *Applied Linear Regression Models*. McGraw-Hill/Irwin, New York, 4 edition, 2004. ISBN 9780073014661.
- [74] Geoffrey Keppel and Thomas D. Wickens. *Design and Analysis: A Researcher's Handbook*. Pearson, Upper Saddle River, NJ, 4 edition, 2004. ISBN 9780130415292.

- [75] Gerard J. P. Van Breukelen. Ancova versus change from baseline: More power in randomized studies, more bias in observational studies. *Journal of Clinical Epidemiology*, 59(9):920–925, 2006. doi: 10.1016/j.jclinepi.2006.02.007. URL <https://doi.org/10.1016/j.jclinepi.2006.02.007>.
- [76] Samuel S. Shapiro and Martin B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 1965. doi: 10.2307/2333709. URL <https://doi.org/10.2307/2333709>.
- [77] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2):486–489, 2012. doi: 10.5812/ijem.3505. URL <https://doi.org/10.5812/ijem.3505>.
- [78] Graeme D. Ruxton. The unequal variance t-test is an underused alternative to student’s t-test and the mann–whitney u test. *Behavioral Ecology*, 17(4):688–690, 2006. doi: 10.1093/beheco/ark016. URL <https://doi.org/10.1093/beheco/ark016>.
- [79] Morton B. Brown and Alan B. Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974. doi: 10.1080/01621459.1974.10482955. URL <https://doi.org/10.1080/01621459.1974.10482955>.
- [80] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. Wiley, Hoboken, NJ, 3 edition, 2014. ISBN 9780470387375.
- [81] Hyun-Yong Kim. Analysis of variance (anova) comparing means of more than two groups. *Restorative Dentistry & Endodontics*, 39(1):74–77, 2014. doi: 10.5395/rde.2014.39.1.74. URL <https://doi.org/10.5395/rde.2014.39.1.74>.
- [82] Douglas C. Montgomery. *Design and Analysis of Experiments*. Wiley, Hoboken, NJ, 10 edition, 2019. ISBN 9781119492442.

- [83] Andrew J. Vickers and Douglas G. Altman. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ*, 323(7321):1123–1124, 2001. doi: 10.1136/bmj.323.7321.1123. URL <https://doi.org/10.1136/bmj.323.7321.1123>.
- [84] Douglas G Altman and J Martin Bland. Standard deviations and standard errors. *BMJ*, 331:903, 2005. doi: 10.1136/bmj.331.7521.903.
- [85] A. Ralph Henderson. The bootstrap: A technique for data-driven statistics. using computer-intensive analyses to explore experimental data. *Clinica Chimica Acta*, 359: 1–26, 2005. doi: 10.1016/j.cccn.2005.03.002.
- [86] Douglas G Altman and J Martin Bland. Parametric v non-parametric methods for data analysis. *BMJ*, 338:a3167, 2009. doi: 10.1136/bmj.a3167.
- [87] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [88] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.
- [89] Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, 1996.
- [90] Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer, 1995.
- [91] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, 1997.
- [92] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, 2000.
- [93] David A. Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9(6): 1218–1228, 1981.

- [94] C. F. Jeff Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295, 1986.
- [95] Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.
- [96] Andrew P. Grieve. Extending a bayesian analysis of the two-period crossover to allow for baseline measurements. *Statistics in Medicine*, 13:905–929, 1994.
- [97] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
- [98] Steven N Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999. doi: 10.7326/0003-4819-130-12-199906150-00008.
- [99] Richard Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall/CRC Press, Boca Raton, FL, 1997.
- [100] Peter C Austin, Leslie J Brunner, and Janet E Hux. Bayeswatch: an overview of bayesian statistics. *Journal of Evaluation in Clinical Practice*, 8(2):277–286, 2002. doi: 10.1046/j.1365-2753.2002.00338.x.
- [101] Simon French and J.Q. Smith. Bayesian analysis. In Simon French and J.Q. Smith, editors, *The Practice of Bayesian Analysis*, pages 1–24. Arnold, London, 1997.
- [102] Leonard Freedman. Bayesian statistical methods. *British Medical Journal*, 313:569–570, 1996. Editorial.
- [103] Steven N Goodman. Toward evidence-based medical statistics ii: the bayes factor. *Annals of Internal Medicine*, 130(12):1005–1013, 1999.

- [104] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Introducing markov chain monte carlo. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 1–19. Chapman & Hall, London, 1996.
- [105] Donald A Berry. A case for bayesianism in clinical trials. *Statistics in Medicine*, 12: 1377–1393, 1993.
- [106] Cheryl L Christiansen and Carl N Morris. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine*, 127(9):764–768, 1997.
- [107] James M Brophy and L Joseph. Practice variations, chance and quality of care. *Canadian Medical Association Journal*, 159:949–952, 1998.
- [108] Dalene K Stangl and Donald A Berry. *Meta-Analysis in Medicine and Health Policy*. Marcel Dekker Inc., New York, 2000.
- [109] Andrew R Willan and Joseph L Pater. Using baseline measurements in the two-period crossover clinical trial. *Controlled Clinical Trials*, 7(4):282–289, 1986. doi: 10.1016/0197-2456(86)90036-x.
- [110] A. O. Varma and N. W. Chilton. Crossover designs involving two treatments. *Journal of Periodontal Research. Supplement*, 14:160–170, 1974. doi: 10.1111/j.1600-0765.1974.tb01780.x.
- [111] Michael G Kenward and Byron Jones. The analysis of data from 2 x 2 cross-over trials with baseline measurements. *Statistics in Medicine*, 6(8):911–926, 1987. doi: 10.1002/sim.4780060806.
- [112] Eric M. Chi. Analysis of cross-over trials when within-subject errors follow an ar(1) process. *Biometrical Journal*, 34:359–365, 1993.
- [113] H. I. Patel. Use of baseline measurements in the two-period crossover design. *Communications in Statistics - Theory and Methodology*, 12:2693–2712, 1983.

- [114] Seymour Geisser. Estimation in the uniform covariance case. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26:477–483, 1964.
- [115] Rens van de Schoot, David Kaplan, Jaap Denissen, Jens B Asendorpf, Franz J Neyer, and Marcel AG van Aken. A gentle introduction to bayesian analysis: applications to developmental research. *Child Development*, 85(3):842–860, 2014. doi: 10.1111/cdev.12169.
- [116] Andrew P Grieve. A bayesian analysis of the two-period crossover design in clinical trials. *Biometrics*, 41:979–990, 1985. Correction: *Biometrics*, 42, 459 (1986).
- [117] Harold Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935. doi: 10.1017/S030500410001330X.
- [118] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.
- [119] Susan M. Lewis and Adrian E. Raftery. Estimating bayes factors via posterior simulation with the laplace–metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.
- [120] James O. Berger. The case for objective bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. doi: 10.1214/06-BA115.
- [121] I. J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. Dover Publications, Mineola, NY, 2 edition, 2009.
- [122] Jose M. Bernardo and Rafael Rueda. Bayesian hypothesis testing: A reference approach. *International Statistical Review / Revue Internationale de Statistique*, 70(3):351–372, 2002.

- [123] Michael D. Lee and Eric-Jan Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Cambridge, 2014.
- [124] D. J. Spiegelhalter and A. F. M. Smith. Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44:377–387, 1982.
- [125] I. J. Good. *Probability and the Weighing of Evidence*. Charles Griffin and Company, London, 1947.