

Faculté des sciences

Generalized pairwise comparison applied to rare diseases trials

Author: **Kamal Gasser**

Supervisors: **Prof. Catherine Legrand, Dr. Mickaël De Backer**

Reader: **Dr. Marc Buyse**

Academic year 2023–2024

Master [120] en statistique, orientation biostatistiques

Acknowledgements

I would first like to thank my friend L  na  c Frehner, who has always been there for me and encouraged me to return to school. I also want to express my gratitude to my mother, who has consistently supported me regardless of the situation and has always been proud of me. The Fondation Huber Tuor deserves special thanks as well for supporting me throughout the two years of this master's program; without their support, this achievement would have been impossible.

I am deeply grateful to my supervisors, Prof. Catherine Legrand and Dr. Mickael de Backer. They provided me with the opportunity to work on a truly interesting subject, offered invaluable feedback and advice, and helped me navigate through challenges. Their thorough review and corrections of my documents have greatly improved the quality of my work.

Additionally, I would like to thank Dr. Marc Buyse for serving on my Master's committee. I greatly appreciate the valuable time he has taken to review this work and his insightful contributions.

Abstract

Clinical trials for rare diseases face significant challenges, including limited patient populations, variability in symptoms, and the complexity of trial designs such as those measuring outcomes at baseline versus the end of the treatment period without a control group (Kandi and Vadakedath 2023). The Generalized Pairwise Comparison (GPC) methodology offers a potential solution by allowing the inclusion of multiple endpoints in the analysis, thereby eliminating the need to select a single primary endpoint (Buyse 2010). In this thesis, we employ linear mixed models to simulate the data and evaluate the performance of the GPC approach, focusing on a comparison between the matched and unmatched methodologies.

The two methodologies measure different aspects of treatment effects, the matched method measure Γ , the proportion of individuals who show improvement under treatment compared to their baseline. While the unmatched method measure Δ , the expected improvement of a patient across the entire population when is under treatment rather than the standard care. Unless the two groups are nearly independent, both methods exhibit similar statistical power. However, due to the interpretability of Γ , we advise to use the matched method in a baseline versus end-of-treatment design, because correlation between the two measures is expected.

Applying these methods to real data from the Sirolimus trial, we find a strong concordance between the two approaches, further validating their use. While GPC shows great promise as a tool for analyzing rare disease trials, one major limitation is its current inability to account for the longitudinal nature of the data. Future research should aim to extend the GPC methodology to incorporate longitudinal data, thereby enhancing its utility in clinical trials.

Contents

Acknowledgements	1
Abstract	2
1 Introduction	9
2 Clinical trial design	11
2.1 Introduction	11
2.2 parallel groups design in clinical trials	12
2.2.1 Variance of the estimated treatment effect	13
2.2.2 Limitations and challenges	13
2.3 Crossover design	13
2.3.1 Increasing precision	14
2.3.2 Carryover effects	14
2.3.3 Dropout issues	15
2.3.4 Considerations for using crossover designs	15
2.4 Rare diseases	15
2.4.1 Recruitment challenges	15
2.4.2 Selection of primary endpoint	15
2.4.3 Statistical power and analysis	16
3 Theoretical foundations of generalized pairwise comparison	18
3.1 Introduction	18
3.2 Measure of treatment effect in GPC	19
3.3 Overview of generalized pairwise comparison methods	21
3.4 Estimation of net benefit	22
3.4.1 Unmatched net benefit	22
3.4.2 Matched net benefit	22
3.5 Inference on net benefit	23
3.5.1 Unmatched net benefit	23
3.5.1.1 Randomization test	23

3.5.1.2	Asymptotic test	23
3.5.1.3	Studentized permutation test in paired data	24
3.5.2	Matched net benefit	26
4	Linear mixed model	28
4.1	Introduction	28
4.2	Linear mixed effects model	28
5	Simulation matched vs. unmatched GPC	30
5.1	Introduction	30
5.2	Single outcome	30
5.2.1	Sampling distribution	30
5.2.2	Theoretical net benefit	33
5.2.2.1	Unmatched net benefit	33
5.2.2.2	Matched net benefit	33
5.2.3	Power	34
5.2.3.1	The naive approach	34
5.2.3.2	Results for the comparison of the matched vs unmatched GPC	36
5.2.4	Extreme scenario	40
5.3	Two continuous outcomes	42
5.3.1	Simulation setup	42
5.3.2	Results for the comparison of the matched vs. unmatched GPC . . .	43
5.4	Discussion	53
6	Application to the sirolimus trial data	54
6.1	Introduction	54
6.2	Classification of various anomalies	55
6.2.1	Lymphatic Malformations	55
6.2.2	Venous Malformation	55
6.2.2.1	Maffucci's syndrome	55
6.2.2.2	Blue Rubber Bleb Nevus Syndrome of Bean	56
6.2.3	Capillary Malformations	56
6.2.4	Conclusion	56
6.3	Analysis of the data	57
6.3.1	Exploratory analysis	58
6.3.1.1	Overall Results	60
6.3.1.2	Group-Specific Findings	60
6.3.2	Linear-mixed effect model	60
6.3.3	Generalized pairwise comparison	62
6.3.3.1	Matched vs. unmatched GPC Analysis of the sirolimus Trial	63
6.3.4	Application Shiny	64
7	Conclusion	67
A	Simulation plan for the two continuous outcome	69
B	Bar plot of various variables from the Sirolimus trial	70

List of Figures

5.1	Illustration of one dataset generated for the simulations for different DGP and the data from the sirolimus trial each with 132 patients. On the top left one with low variability and no correlation between the random effect, on the top right one with high variability and no correlation between the random effect while on the bottom left one with high variability and high correlation between the random effect. Finally, on the bottom right the data collected during the sirolimus trial	32
5.2	Monitoring of type I error for the matched and unmatched GPC under various values of the variance parameters	35
5.3	Monitoring of type I error for the matched and unmatched GPC under various values of the variance parameters when permuting only at the individual level	36
5.4	Comparison of the net benefit for the matched GPC vs. unmatched GPC for different values of the variance parameters when the sample size is set to 10, the bar represents the mean +/- one standard deviation	37
5.5	Comparison of the net benefit for the matched GPC vs. unmatched GPC for different values of the variance parameters when the sample size is set to 100, the bar represents the mean +/- one standard deviation	38
5.6	Type I error simulations of unmatched GPC vs. matched GPC inference for different values of the variance parameters when the true treatment effect (λ) is set to 0 for testing hypothesis of the kind $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ and $H_0 : \Gamma = 0$ vs. $H_1 : \Gamma \neq 0$. The number on the bottom left of each panel represents the correlation between the two the two measures (baseline/end of treatment period) induced by the random effect	39
5.7	Power comparison of unmatched GPC vs. matched GPC for different values of the variance parameters when the true treatment effect (λ) is to -1 for testing hypothesis of the kind $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ and $H_0 : \Gamma = 0$ vs. $H_1 : \Gamma \neq 0$. The number on the bottom left of each panel represents the correlation between the two measures (baseline/end of treatment period) induced by the random effect	40
5.8	Power comparison of the matched GPC versus unmatched GPC for extreme settings is shown in Figure 5.8. On the left (1), the correlation between the endpoints is near 0, while on the right (2), we have a correlation between the endpoints of 0.996.	41
5.9	Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on both outcomes, meaning that we used the values of the parameters from extreme scenario 1 (see Section 5.2.4) with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 1 in Appendix A).	44

5.10	Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on both outcomes, meaning that we used the values of the parameters from extreme scenario 1 (see Section 5.2.4). With positive correlation between the two outcomes at the end of the treatment period (see simulation plan setup 2 in Appendix A).	45
5.11	Both simulations have the same setup, namely the parameters form extreme scenario 1 (see Section 5.2.4) favoring the unmatched methods for both outcomes with a treatment effect of -1 on both outcomes but with the amount of correlation between outcome 1 and outcome 2 at the end of the treatment period set to -0.5 on top and to 0.5 on the bottom. (see simulation setup 1 and 2 in the Appendix A)	46
5.12	Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on outcome 1 and the matched GPC on outcome 2 with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 3 in Appendix A)	48
5.13	Comparison of the matched vs. unmatched GPC when we try to favor the matched methods on outcome 1 with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 5 in Appendix A)	49
5.14	Comparison of the matched vs. unmatched GPC when we favor the matched methods on both outcomes (using the values of parameters form extreme scenario 2 from Section 5.2.4) with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 7 in Appendix A)	50
5.15	Comparison of the matched vs. unmatched GPC when the measures at baseline and at the end of treatment are highly correlated, using the values of parameters form extreme scenario 2 from Section 5.2.4 for both outcomes. The measures at baseline of outcome 1 and outcome 2 are highly correlated (see simulation plan setup 9 in Appendix A).	51
5.16	Comparison of the matched vs. unmatched GPC when the measures at baseline and at the end of treatment are highly correlated, using the values of parameters form extreme scenario 2 from Section 5.2.4 for both outcomes. The measures at baseline of outcome 1 and outcome 2 are highly correlated (see simulation plan setup 9 in Appendix A).	52
6.1	The left bar plot displays the proportion of individuals reporting pain at each follow-up. The right bar plot shows the proportion of individuals reporting pain, categorized by their diagnostic groups.	58
6.2	The left bar plot displays the proportion of individuals reporting pain exacerbations at each follow-up. The right bar plot shows the proportion of individuals reporting pain exacerbations, categorized by their diagnostic groups.	59
6.3	The left bar plot displays the proportion of individuals reporting functional limitations at each follow-up. The right bar plot shows the proportion of individuals reporting functional limitations, categorized by their diagnostic groups.	59

6.4	Visual representation of the linear mixed model for pain frequency over time (months). Individual participant trajectories (colored lines) are shown alongside the population average trend (blue dashed line) and its 95% confidence interval (grey shaded area).	62
6.5	Comparison of the distribution of pairs for the matched vs. unmatched GPC analysis of the sirolimus trial prioritizing functional limitation, functional limitation frequency, and functional limitation severity as outcomes.	63
6.6	Tab 'Load the data' of the sirolimus Shiny application	65
6.7	Tab 'Linear mixed model' of the sirolimus Shiny application	65
6.8	Tab 'GPC' of the sirolimus Shiny application	66
B.1	The left bar plot displays the mean pain severity reported at each follow-up with +/-standard deviation. The right bar plot shows mean pain severity reported, categorized by their diagnostic groups.	71
B.2	The left bar plot displays the mean pain frequency reported at each follow-up with +/-standard deviation. The right bar plot shows mean pain frequency reported, categorized by their diagnostic groups.	71
B.3	The left bar plot displays the mean functional limitation frequency reported at each follow-up with +/-standard deviation. The right bar plot shows mean functional limitation frequency reported, categorized by their diagnostic groups.	72
B.4	The left bar plot displays the mean functional limitation frequency reported at each follow-up with +/-standard deviation. The right bar plot shows mean functional limitation frequency reported, categorized by their diagnostic groups.	72
B.5	The left bar plot displays the mean pain exacerbations frequency reported at each follow-up with +/-standard deviation. The right bar plot shows mean pain exacerbations frequency reported, categorized by their diagnostic groups.	73
B.6	The left bar plot displays the mean pain exacerbations severity reported at each follow-up with +/-standard deviation. The right bar plot shows mean pain exacerbations severity reported, categorized by their diagnostic groups.	73

List of Tables

3.1	Possible outcomes for generalized pairwise comparisons according to the type of variable	21
5.1	Empirical vs. theoretical derivation of the the net benefit for the matched and unmatched GPC for various configuration.	38
5.2	Value of the various parameters for the extreme case, the first scenario corresponds to the one where the two measures (baseline/end of treatment period) are not correlated at all, while the second setting correspond to a really high amount of correlation between the two measures (0.996).	41
6.1	Estimated fixed effects and variance components for the random intercept and random slope model of frequency of pain, with time (in months) as the covariate	61
6.2	Estimated fixed effects and variance components for the random intercept and random slope model of functional limitation severity, with time (in months) as the covariate	61
A.1	Simulation plan for for the two continuous outcome.	69

1. Introduction

The main goal of this thesis is to investigate the implementation of the Generalized pairwise comparison (GPC) methodology within the context of rare diseases trials, where typically a control group is lacking, and where the selection of a unique primary clinical endpoint for analysis might be difficult because of the complex nature of rare diseases (Rath et al. 2017). Rare disease is defined as one affecting fewer than 200,000 individuals in the US [prevalence, about six per 10,000] and up to five per 10,000 in Europe [affecting about 250,000]. Individual rare diseases are, for the most part, rare, but collectively, they constitute some 7,000 different named diseases. About 80 per cent of rare diseases are genetic, often devastating and relentless. A major, and obvious, limitation of rare disease clinical trials is the small number of patients available for recruitment. On top of that, the selection of correct, relevant and meaningful endpoints for rare disease clinical trials requires a vastly enhanced knowledge of the natural history of the disease. However, rare diseases are heterogeneous, and often obscure, with little information available on their natural history; in addition, expression of these diseases and their subtypes is often variable and, frequently, related to specific genotypes. Most rare diseases are complex, affecting multiple systems making endpoint selection even more challenging (Mellerio 2022).

The GPC methodology is one possible solution because it allows the incorporation of multiple, hierarchically ordered endpoints, giving an overall picture about the efficacy of a treatment. In this work we focused on the efficacy assessment of the GPC in the situation where all the patients are treated with the ‘experimental treatment’ (while their own baseline measurements are used as control). This setup is motivated by the sirolimus trial, a phase III clinical trial aimed at assessing the efficacy of Sirolimus in improving the quality of life of patients diagnosed with slow-flow vascular malformations. As discussed in Section 6, there are various forms of slow-flow vascular malformations, each with a wide range of symptoms, even within the same subtype. In this context, determining the primary endpoint is not straightforward, making the GPC methodology a suitable tool to assess the treatment effect.

Crucial to this work is the comparison between the matched and unmatched GPC analyses used in this trial design. The matched method will compare for each individual patient the outcomes at baseline and at the end of the treatment period (effectively treating each patient as their own control). The unmatched method will treat the baseline measurements and the end of treatment measurements as if they are two separate groups and perform pairwise comparisons between those groups. It is often argued that the unmatched method will be more powerful as it performs all possible pairwise comparisons between the baseline and the treatment period. However, as the simulation study (Section 5) demonstrated, this is not always the case. A key aspect of this work is to realise that the two methods measure distinct parameters. The matched net benefit (Γ) is the proportion of individuals who do better under treatment than at baseline. This metric is easily interpretable and communicable. The unmatched net benefit (Δ) measures the expected improvement across the population in term of quantiles for a patient who has received the treatment relative to baseline. One factor

influencing why one method has superior power over the other is the correlation between baseline and end of treatment measurements, as well as the correlation between outcomes.

The thesis is structured as follows: Section 2 provides an overview of clinical trial designs, encompassing parallel group designs, crossover designs, and considerations specific to rare disease trials. Section 3 establishes the theoretical foundation of the GPC methodology, including measures of treatment effect, estimation of net benefit, and inference procedures. Section 4 introduces linear mixed models, which are employed for data generation and analysis in subsequent sections. Section 5 presents a simulation study comparing the power of matched and unmatched GPC approaches under various scenarios. Finally, Section 6 applies the GPC methodology to the sirolimus trial data, offering insights into its practical implementation and interpretation in the context of rare disease research.

2. Clinical trial design

2.1 Introduction

Clinical trials are structured experiments that evaluate the safety and efficacy of new treatments in humans; these trials determine whether an experimental novel treatment is sufficiently safe and effective to be applied broadly to patient populations. Due to individual variability, controlled research is necessary to determine the efficacy of a therapy objectively. Randomised controlled trials (RCTs) are the gold standard for quantitative evidence in healthcare. A RCT compares the effects of an experimental, novel intervention with what we called a control in this thesis, which might be a standard of care, a placebo, another comparator therapy of the same condition or the condition under no treatment in the case of chronic diseases. Without RCT the assessment of the safety and benefits of a novel medical breakthrough might be compromised, and patients could potentially be exposed to medicines that are not only useless but also potentially harmful (Kandi and Vadakedath 2023). The primary endpoint is a central concept in clinical trial design, the practitioner must define the primary endpoint which is the main outcome measure of the trial. The primary endpoint is the main indicator of efficacy or effectiveness of the intervention being studied. A defined main endpoint offers quantifiable focus for the study and is required for statistical evaluation and interpretation. In defining the primary endpoint, consideration must be given to a series of factors, including:

- The nature of the illness or condition that is investigated: this will inform what outcome measures are the most important.
- The principal endpoint should match the most important goal of the trial.
- Is the endpoint feasible to measure in the context of the trial and among the target population ?

Ultimately, choice of the primary endpoint remains a difficult decision that should be carefully considered and is a key determinant of success or failure of a clinical trial (Kandi and Vadakedath 2023).

There is a broad literature on clinical trials involving a variety of trial procedures and statistical methodologies. However, the aim of this section is not to discuss this vast topic in details but to describe two of the standard study designs used in clinical trials, the parallel groups design and the cross-over design, we will also discuss some concern about clinical trials for rare diseases. The parallel groups design consists of two groups, where one group receives the drug being tested (treatment group) and the second group gets the control (the control group). The participants are assigned to each group randomly. The control could be a placebo, no drug at all or it could be a competitor drug. The features of parallel groups designs are as follows:

- Two treatment groups, one receiving the treatment and the other the control. Each of the patients is randomly assigned to one of these.
- Both groups receive their respective treatments simultaneously.
- Subjects in each group are independent.
- Statistical analysis compares outcomes between the two arms of the study to deduce the efficacy of the treatment.

The second is the so-called cross-over design where each subject receives both the drug under investigation and the ‘control’, i.e., in that particular study each individual serves as their own control. In the simplest case, each subject is randomly assigned to receive one of the treatments – we label it A – in the first period and the complementary treatment – let’s say B – in the following period. The features of cross-over design are as follows:

- Each subject experiences both the experimental treatment and the control, removing the between subject variability.
- The order in which the participants are given the treatments (e.g, AB or BA) is randomly determined.
- There is often a gap of time between the treatment periods called the ‘washout period’ to ensure that the effects of the first drug have dissipated entirely before the second treatment is administered.
- The statistical analysis compares the individual’s response with treatment A to that with treatment B.

This is an attractive design for researchers, because it provides the opportunity to test the effect of each of the treatments on the same participant, which potentially minimises interindividual variability and allows for smaller sample sizes (Friedman 2015).

The goal of this thesis is to evaluate how the GPC methodology can be applied to rare diseases where, for each patient, we have outcome values both at baseline and at the end of the treatment period. If we consider the baseline values as the control group and the end of treatment values as another group, we arrive at a setup similar to a parallel-group design. However, if we assess the treatment effect by comparing the baseline versus end-of-treatment values for each patient individually, we approach a setting similar to a crossover design.

In Section ??, we discuss the parallel groups design, followed by the crossover design in Section 2.3. Finally, in Section 2.4, we address the main characteristics of rare disease clinical trials.

2.2 parallel groups design in clinical trials

parallel groups design is the most commonly used design in clinical trials, and more specifically randomised controlled trials (RCTs). The main reason for adopting this design is that it is simple, robust, efficient and suitable for any intervention and almost any outcome. In this design, participants are randomly allocated to two or more parallel groups, with each group receiving one of the interventions simultaneously. The comparison of the outcomes of these groups is carried out under similar conditions, which aims to minimise bias and make the results comparable.

The key feature of a parallel groups design is randomising participants to either the treatment group (who receive the experimental intervention), or the control group (who receive another standard treatment, placebo or no treatment). Randomisation helps prevent other confounding variables from explaining any differences between groups. The major strength of the parallel groups design is its simplicity. Because all groups are treated at the same time, it avoids the temporal bias that can affect crossover studies, where each research subject receives interventions in a sequential manner. Moreover, the parallel design allows data to be collected at the same points in time from each arm of the study, which results in a simple analysis and interpretation of results.

2.2.1 Variance of the estimated treatment effect

In a parallel groups design with two treatments (call them A and B) where each group has N patients enrolled, the estimate of the treatment effect can be given by $\hat{\Phi} = \bar{Y}_A - \bar{Y}_B$. Where the $\bar{Y}_i$ ($i \in \{A, B\}$) is the mean outcomes for treatment A and B respectively. The variance of this estimator ($V(\hat{\Phi})$) is very important since it gives an idea of the precision and we will see in the next section how we can use it to compare the precision obtained in other designs. If the variance of the outcome measure is the same in both groups (σ^2) and design balanced with N participant in both groups, it is not difficult to show that the variance of the estimator of the difference of the treatment effect is:

$$V(\hat{\Phi}) = 2\frac{\sigma^2}{N}. \quad (2.1)$$

We can read off from this expression that, as the size of the sample increases, the variance of the estimator shrinks. The smaller the variance is, the smaller the sample size need be to achieve a given power.

2.2.2 Limitations and challenges

Although the parallel groups design has many advantages, it also has limitations and challenges. One of the main limitations is the risk of baseline imbalances between groups, particularly in smaller trials, which can lead to biased estimates of treatment effects. Strategies such as stratified randomization and minimization reduce this risk, but can make the randomization more complex. Another problem is the need for large numbers of trial participants to reach adequate statistical power, especially if the predicted treatment effects are small. This can make parallel groups trials very expensive and time-consuming. Furthermore, since the treatment allocation is fixed at the start, once a participant has been assigned to a group, they remain in that group for the duration of the study, which can lead to problems with compliance and dropout.

2.3 Crossover design

In contrast to the parallel groups design, a crossover trial involves each patient receiving every treatment but at different times. This design allows for the study of treatment effects by comparing the outcomes within the same patient, thereby serving as their own control. A significant concern in crossover trials is the potential for carryover effects, where a previous treatment can influence the outcome of a subsequent treatment. The simplest crossover design includes two treatments, denoted as A and B, and two time periods. In this setup,

subjects are randomly assigned to receive treatment A first and then treatment B, or vice versa. This design, referred to as AB/BA, will be the focus of this discussion. However, it should be noted that more complex designs exist, such as those with more than two periods for each treatment or incomplete block designs.

Randomization in a crossover trial differs from that in a parallel groups design. Since each patient receives all treatments, only the order in which treatments are administered is randomized. Therefore, randomization alone does not ensure an unbiased comparison of treatments; the validity of the comparison relies on the assumption that treatment effects are confined to the period of administration.

2.3.1 Increasing precision

One of the primary advantages of crossover trials is the increased precision in estimating treatment effects. Since each patient serves as their own control, within-subject variability is reduced. This reduction in variability decreases the required sample size needed to achieve the same precision in estimating the treatment effect, ultimately reducing the number of subjects needed to achieve similar power as in a parallel groups design.

The reduction in the number of subjects is two-fold. First, because each subject receives both treatments, only half the number of subjects is needed. Second, the within-patient responses to the treatments are often positively correlated, further reducing the variability of the estimated treatment effect and increasing precision.

For example, consider the treatment effect in an AB/BA design with a constant variance σ^2 in each patient for each treatment, assuming no carryover effect. An estimator of the difference in treatment effect is given by $\hat{\Phi} = \bar{Y}_A - \bar{Y}_B$, where $\bar{Y}_A$ and $\bar{Y}_B$ are the average responses on treatments A and B, respectively. The variance of this estimator is given by:

$$\mathbb{V}(\hat{\Phi}) = \frac{\sigma^2}{N} + \frac{\sigma^2}{N} - 2\text{cov}(\bar{Y}_A, \bar{Y}_B) = 2\frac{\sigma^2}{N}(1 - \rho_{AB}). \quad (2.2)$$

Where ρ_{AB} is the within-subject correlation of responses on treatments A and B, assumed to be the same for all individuals, and N is the sample size. If ρ_{AB} is greater than zero, there is a reduction in the variability of the estimator compared to a parallel groups design (see Equation 2.1). Conversely, if this correlation is negative and large (though unlikely), the crossover design would be less efficient than a parallel design.

2.3.2 Carryover effects

Carryover effects occur when the treatment effect from one period continues into the following period. This can happen if:

- The treatment agent remains physiologically active during the second period, which can be prevented by allowing a sufficiently long washout period between treatments.
- The treatment causes a permanent change or cures the patient, leading to biased estimates. For example, if the first treatment cures the patient, the second treatment may appear more effective than it is.
- The patient's condition changes over time, resulting in a true interaction between the period and the treatment effect.

Analytic approaches exist to test and address carryover effects, but this thesis does not cover them in details. For a comprehensive discussion, refer to Senn (2002). The simplest way to manage carryover effects is to incorporate a sufficiently long washout period and assume no carryover effect. The appropriate length of the washout period is often determined by the half-life of the pharmaceutical product within the population of interest, typically using a multiple of the half-life.

2.3.3 Dropout issues

Dropout rates can be more severe in crossover trials than in parallel designs. The longer duration of crossover trials increases the likelihood of patients leaving the study. Additionally, the exposure to all treatments raises the chances of side effects, leading to higher dropout rates.

2.3.4 Considerations for using crossover designs

While crossover trials offer benefits over parallel designs, certain considerations must be addressed. First, there should be some understanding of the sign and magnitude of the within-patient correlation between responses. A large negative correlation would make the crossover design a poor choice. Second, the disease under investigation should have constant intensity. Crossover designs are unsuitable for diseases that might be cured or for patients who may die during the trial. They are best suited for chronic diseases. Third, a sufficiently long washout period is crucial to avoid simultaneous effects of both treatments.

2.4 Rare diseases

Because of the low prevalence of people living with rare diseases, which are defined variously as those affecting fewer than 200,000 individuals in the US or fewer than 1 in 2,000 in Europe, aspects of trial design such as recruitment, selection of valid and measurable endpoints and statistical power are significantly different from those in other disease areas. This section highlights the specific features of rare disease trials, in terms of recruitment difficulties, endpoint options and statistical power as reported by Rath et al. (2017).

2.4.1 Recruitment challenges

One of the biggest challenges of designing clinical trials for rare illnesses is patient recruitment, as successful sample sizes may necessitate collaboration across several centers, sometimes even across national boundaries. This multi-center approach has its own set of difficulties, including coordinating the protocol design and managing ethical review, as well as dealing with language challenges and diverse legal requirements for the local populations participating. Additionally, because patients with rare diseases frequently have exceptional clinical burdens, recruiting them might be difficult due to the stress of clinical trials and the need for long travel, which can be difficult for patients or their families.

2.4.2 Selection of primary endpoint

The disease heterogeneity and complexity of rare diseases make it difficult to define appropriate primary endpoints for clinical studies. In order to find endpoints that are both

clinically meaningful as well as feasible to measure, it is important to have an understanding of the natural history of the disease and its clinical variability. Standard endpoints used for common diseases are often not suitable for rare diseases due to different disease mechanisms and patient populations (Mellerio 2022; Hilgers 2016).

Rare disease groups tend to have considerable clinical heterogeneity, including variation in age at onset, duration of disease, severity of illness, rate of progression, response to treatment, health-related quality of life, functional status and other outcomes such as productivity, psychological impact, social functioning. Heterogeneity raises the risk that an intervention may have a modest effect on each subgroup, none of which is statistically significant while a combination of those patients would benefit from the intervention. RCT designs typically focus on estimating the average treatment effect for a chosen subgroup of patients, which does not fully account for this heterogeneity. Regulatory agencies such as the FDA and EMA have authorized treatments for rare diseases based on the possibility of clinical benefit, as indicated through certain surrogate endpoints, recognizing the promise of more concrete evidence expected from post-marketing studies. Such adaptive pathways promise to accelerate access to treatments. But these pathways can also be dangerous, potentially permitting ineffective or harmful interventions to enter the market. Patients and decision-makers alike are now demanding patient-centered, clinically meaningful outcomes.

Patient-centered outcomes, which measure the impact of a therapy on patients' life, daily functioning or overall well-being, can be clinician-reported, observer-reported or patient-reported, based on patient-reported outcome measures. As the clinical presentation of rare diseases is generally more complex, a singular clinical outcome is often not enough to characterize the impact of a treatment. To achieve statistically based comprehensive clinical efficacy assessment, multiple clinical outcomes will have to be considered. This increases the complexity of the statistical evaluation. Novel methodologies such as Generalized Pairwise Comparisons might be particularly useful in this context, by allowing a hierarchical specification of multiple outcomes which enable a deeper understanding of treatment effects and can substantiate the success of small trials with promising new treatments in the context of a rare disease population.

2.4.3 Statistical power and analysis

Given the inherent small sample sizes in rare disease trials, powering trials to detect an effect is often challenging. Standard frequentist methods designed for larger sample sizes often fail to achieve statistical power. Exploratory and innovative approaches, like adaptive trial design and innovative Bayesian methods, are evolving to increase the statistical efficiency of rare disease trials. We refer the reader to the review made by Abrahamyan et al. (2016) for a more detailed treatment of the different rare-disease designs. For the needs of this thesis, one design of interest for which we could not find any name in the literature is the following: baseline versus (vs.) end of treatment period. This type of design is only appropriate when the disease is chronic and remains stable without significant positive or negative progression. Otherwise, the natural evolution of the disease and the treatment effect become confounded. In this design, every patient is treated with the new drug and the control arm is the value at baseline. The treatment effect is evaluated by the comparison between the value of outcomes at baseline vs the value of outcomes at end of the period treatment. The rest of this thesis will be devoted at exploring how the GPC methodology can be applied to the following design. We will make the distinction between two types of analysis, the matched analysis, in this setting we will compare the value of the outcomes at the individual level and each

patient will serve as his own control similarly as in the cross-over design. For the second type of analysis that we will call the unmatched analysis, we will imagine that we have two groups, the control group which will correspond to the values observed at baseline and the treatment group which will correspond to values observed at the end of the treatment period. As we will see in Section 5 even though we are making the experimental thought that we have two distinct group, we cannot assume that the groups are independent as the value at baseline and at the end of the treatment period for an individual is expected to be correlated.

3. Theoretical foundations of generalized pairwise comparison

3.1 Introduction

The objective of this work is to use the Generalized prioritized pairwise comparison methods (GPC) in the context of rare diseases trial. In particular when we do not have a control arm but we have values of the outcomes at baseline and at the end of the treatment period and when it's difficult to choose a primary outcome of interest. We will illustrate this for the special case of the sirolimus trial (Seront et al. 2023) where the goal was to evaluate the efficacy of sirolimus to reduce the symptoms related to slow-flow vascular malformation (see section 6).

Suppose we have N patients with outcome values recorded at baseline. Each patient then undergoes treatment, and their outcome values are measured again at the end of the treatment period. We will get the following pairs $(\mathbf{Y}_{i0}, \mathbf{Y}_{i1})'$ of a random vector $\mathbf{Y}_{ik} = (Y_{ik1}, \dots, Y_{ikL})'$ where i denote the patient index ($i \in \{1, 2, \dots, N\}$), k the group ($k \in \{0 = \text{baseline}, 1 = \text{treatment}\}$) and l the various outcomes ($l \in \{1, 2, \dots, L\}$).

Now with this data, we can either realize our analysis via matched test by comparing if the patient had a better outcome when he was under treatment rather than the control treatment. However, we can use the data as if we had two groups namely the control group and the treatment group and pursue our analysis by comparing the two groups. For the first methods we will have access to only N pairs, but for the second we will have N^2 pairs. The fact that we can use both method rises questions.

1. Do we measure the same thing by using one or the other ?
2. Is one of the method uniformly most powerful than the other ? If not which characteristics should be considered when choosing one or the other ?

The first question has been answered by Fay et al. (2018). For simplicity here suppose that there is only one outcome in consideration. To answer our question we need to introduce the concept of potential outcome and counterfactual. In the case where the treatment is binary (treatment/control), before the trial we have two potential outcomes for an individual i . Namely, $Y_i(0)$ which would be the value of Y that would be observed for patient i if he received the control treatment and $Y_i(1)$ which would be the value of Y that would be observed for patient i if he received the treatment. On the other hand, counterfactual outcomes are the ones that would have been observed, had the treatment been different. If patient i was assigned to treatment, then his counterfactual is $Y_i(0)$.

In a parallel randomized clinical trial design, patient are assigned to one group, either the treatment group or the placebo group. Therefore, before the treatment decision is made, any outcome is a potential outcome while at the end of the study, there is an observed outcome

and a counterfactual outcome which is a missing value. However, in a cross-over trial each patient goes under a phase where they get the placebo and a phase in which they are under treatment, therefore, at the end of the study we have two observed outcome.

With this notation we can distinguish two measures that we will define and explain more precisely in the next section :

$$\Delta = \mathbb{P}(Y_i(1) > Y_j(0)) - \mathbb{P}(Y_i(1) < Y_j(0)) \ (i \neq j) \quad (3.1)$$

and the following quantity :

$$\Gamma = \mathbb{P}(Y_i(1) > Y_i(0)) - \mathbb{P}(Y_i(1) < Y_i(0)). \quad (3.2)$$

We see that Γ measures the probability that an individual has a more favorable potential outcome under treatment than under the control group minus the opposite probability. However, Δ measure the probability that an individual in the treatment group has a more favorable potential outcome than an other individual in the control group. It seems natural that Γ is a better quantity to communicate, as a patient is more interested to know if he had a better chance by taking the treatment rather than the placebo. However, Δ is most often communicated, this is due to the fact that in randomized control trials the joint distribution of $(Y_i(0), Y_i(1))$ can't be observed and Γ is not identifiable. Moreover, if we come back to the interpretation of Δ , it can be shown that :

$$\Delta = 2 \cdot \mathbb{E}[\bar{F}(Y_i(1)) - \bar{F}(Y_i(0))] \quad (3.3)$$

where $\bar{F}(y) = \frac{1}{2}(F_0(y) + F_1(y))$ and F is the cumulative distribution. We notice that the last expression is expressed at the individual level. The interpretation of the preceding quantity is as follows, $\bar{F}(Y_i^1)$ measures the position in terms of quantiles for patient i when he is under the treatment as compared to all individuals, those under treatment and those under control and their counterfactual. Similarly, $\bar{F}(Y_i^0)$ measure the position in terms of quantiles for the same individuals when he is under the control as compared to all individuals of either groups and their counterfactual. So, Δ is the expected improvement in term of quantiles in the full population for an individual when he is under the treatment rather than the control treatment.

Now we understand that there is two metrics and two methodologies that can be applied to our data. Section 3.2 will be dedicated to the various measures of treatment effect. Section 3.3 will develop the framework of generalized pairwise comparison. Then Section 3.4 will show how the estimation of the net benefit can be done. Finally, Section 3.5 will show how to make inference with respect to the unmatched net benefit and the matched net benefit. This will provide us with the required tools to perform simulation studies, see Section 5, in order to evaluate when one method performs better than the other.

3.2 Measure of treatment effect in GPC

This Section is based on **empty citationempty citation**. We start by introducing the probabilistic index (PI) Θ which is the treatment effect measure associated with the Wilcoxon-Mann-Whitney test. In the univariate case, the PI is defined as follows:

$$\Theta = \mathbb{P}(Y_{i1} > Y_{j0}) + \frac{1}{2}\mathbb{P}(Y_{i1} = Y_{j0}). \quad (3.4)$$

This measure has various names in the literature such as relative effect, concordance index or the probability of superior outcome. The interpretation of the PI is as follows, it's the probability that a randomly selected patient from the experimental treatment group has a more favorable response than a randomly-selected patient from the control group plus one half that they have a similar response. The PI ranges from 0 to 1, when the value is above 1/2 it indicates that it is more likely that a patient from the treatment group has a better outcome than a patient from the control arm. If the value is below 1/2 the interpretation is similar but in the other direction. And a value of 1/2 means that a patient in the treatment group is as likely to have a better or a worst outcome compared to a patient in the control group.

The next measure is the net benefit (Δ), which is simply the difference between two probabilities; the probability that a random individual from the treatment (Y_{i1}) group has a better outcome than a random individual from the control (Y_{j0}) group minus the probability that a random individual from the treatment group has a worst outcome than a random individual from the control group :

$$\Delta = \mathbb{P}(Y_{i1} > Y_{j0}) - \mathbb{P}(Y_{i1} < Y_{j0}) \ (i \neq j). \quad (3.1)$$

$\Delta = 1$ means that all individuals in the treatment group have a better outcome than those in the control group. On the other hand, $\Delta = -1$ means that the control group is uniformly better than the treatment group and finally, $\Delta = 0$ means that there is no net difference.

As mentioned in section 3.1, we have a similar measure but at the individual level :

$$\Gamma = \mathbb{P}(Y_{i1} > Y_{i0}) - \mathbb{P}(Y_{i1} < Y_{i0}). \quad (3.2)$$

For the following measures, we will present them only for different individuals. Another common measure of treatment effect is the win ratio (Ψ) which is the ratio of the proportion of wins over the proportion of losses.

$$\Psi = \frac{\mathbb{P}(Y_{1i} > Y_{0j})}{\mathbb{P}(Y_{1i} < Y_{0j})}. \quad (3.5)$$

We can notice that a win ratio of 1 means no treatment effect, while a value bigger than 1 indicates treatment benefit.

Finally, we can define the win odds, which is the ratio of the proportion of wins plus half the proportion of ties over the proportion of losses plus half the proportion of ties.

$$\Lambda = \frac{\mathbb{P}(Y_{1i} > Y_{0j}) + 0.5\mathbb{P}(Y_{1i} = Y_{0j})}{\mathbb{P}(Y_{1i} < Y_{0j}) + 0.5\mathbb{P}(Y_{1i} = Y_{0j})}. \quad (3.6)$$

This is similar to the Mann-Whitney test in case of ties. It allows to take all pairs of patients into account. When there is a large number of ties, it emphasises the similarity between the arms (Brunner, Vandemeulebroecke, and Mütze 2021). It's important to note as well that the Net Benefit can be expressed in terms of the Win odds

$$\Delta = \frac{\Lambda - 1}{\Lambda + 1}. \quad (3.7)$$

3.3 Overview of generalized pairwise comparison methods

According to Buyse (2010), generalized pairwise comparisons of prioritized outcome is an extension of the Wilcoxon Mann–Whitney that allows to compare two groups of individuals in terms of several variables. The setup is as follow, assume that there are two groups, one of n_1 individuals with Y_{i1} representing the value of individual i in treatment and one of n_2 individuals with Y_{j0} representing the value for another individual (possibly the same in the case of matched test) j in the control group. This is a typical situation in randomized clinical trials. The results of these two individuals are compared, and the pair is deemed "favorable" if the individual in the treatment group has a better outcome than the individual in the control group, "unfavorable" if the opposite is true, "neutral" if there is no difference between the two individuals' outcomes, or "uninformative" if it is impossible to determine which of the two individuals has a better outcome. In order to be able to do the comparison, a better outcome must be defined for every pair. For example in the case of binary variables, let $Y_{i1} = 1$ indicates a success and $Y_{i1} = 0$ indicates a failure similarly for Y_{j0} . We can see in table 3.1 the different possible outcomes.

For continuous variables, assume that Y_{i1} and Y_{j0} are now continuous and that the larger Y_{i1} the better the outcome similarly for Y_{j0} . In applied settings, it is custom to set a difference threshold in order to be considered meaningful that we will call τ , the different possible outcomes are presented in Table 3.1. Finally, GPC can be used for other types of variables, the unique requirement is that there exists a way to tell which outcome is better for each possible pair.

Table 3.1: Possible outcomes for generalized pairwise comparisons according to the type of variable

Pairwise comparison for binary variables	
Pairwise comparison	Pair is
$Y_{i1} = 1, Y_{j0} = 0$	favorable
$Y_{i1} = 1, Y_{j0} = 1$	neutral
$Y_{i1} = 0, Y_{j0} = 0$	neutral
$Y_{i1} = 0, Y_{j0} = 1$	unfavorable
Pairwise comparison for continuous variables	
Pairwise comparison	Pair is
$Y_{i1} - Y_{j0} > \tau$	favorable
$ Y_{i1} - Y_{j0} \leq \tau$	neutral
$Y_{i1} - Y_{j0} < -\tau$	unfavorable

With that setting in place we can introduce the algorithm of prioritized generalised pairwise comparisons. First, all endpoints are ranked with the most important one on top. The pairwise comparison between two individuals evaluates first the most important endpoint, if one individual has a favourable outcome, then a score is assigned and the algorithm ends for this pair. However, if the pair is uninformative or neutral for this outcome, we evaluate the pair on the next endpoint. We follow that process until we can declare a winner or that the last endpoint is reach.

3.4 Estimation of net benefit

In this section, we will present how to estimate the net benefit using the unmatched and the matched methods. The section is based on Buyse (2010).

3.4.1 Unmatched net benefit

Let's introduce a pairwise indicator $W_{ij}(l)$ for the pair formed by the $(i^{th} \ i \in \{1, \dots, n_1\})$ individual in the treatment group, the $j^{th} \ (j \in \{1, \dots, n_2\})$ individual in the control group which shows if the pair is favourable, unfavourable or neutral up to the l^{th} endpoint $(l \in \{1, \dots, L\})$.

$$W_{ij}(l) = \begin{cases} 1 & \text{if the pair is favorable} \\ -1 & \text{if the pair is unfavorable} \\ 0 & \text{otherwise} \end{cases} \quad (3.8)$$

The net benefit for the l endpoint can then be estimated by :

$$\hat{\delta}(l) = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} W_{ij}(l)}{n_1 \cdot n_2} \quad (3.9)$$

The cumulative proportion in favour of the treatment up to the l^{th} endpoint is then $\hat{\Delta}(l) = \sum_{i=1}^l \hat{\delta}(i)$. If we finally consider all endpoints (L), we reach the prioritized net benefit estimator :

$$\hat{\Delta} = \sum_{l=1}^L \hat{\delta}(l). \quad (3.10)$$

3.4.2 Matched net benefit

Consider now the case where we compare the pair by individuals meaning that for each individuals we have two measures of the outcome, typically one at baseline and one at the end of the treatment period. This time we introduce a pairwise indicator $W_i(l)$ for the pair formed by the $i^{th} \ (i \in \{1, \dots, N\})$ individual which shows if a pair is favourable, unfavourable or neutral for l^{th} endpoint $(l \in \{1, \dots, L\})$, for individual i when he is under treatment vs. control.

$$W_i(l) = \begin{cases} 1 & \text{if the pair is favorable} \\ -1 & \text{if the pair is unfavorable} \\ 0 & \text{otherwise} \end{cases} \quad (3.11)$$

The net benefit for the l endpoint can then be estimated by :

$$\hat{\gamma}(l) = \frac{\sum_{i=1}^N W_i(l)}{N}. \quad (3.12)$$

Finally if we consider all endpoints (L), we reach the prioritized net benefit estimator :

$$\hat{\Gamma} = \sum_{l=1}^L \hat{\gamma}(l). \quad (3.13)$$

3.5 Inference on net benefit

In this section, we will present how to make inferences about the net benefit for the unmatched and the matched methods. The first two methodologies we present for the unmatched method are those used for classical designs where we have two independent groups. As we will see in Section 5, using these two methods naively in a setting where we have measurements at baseline and at the end of the treatment period leads to uncontrolled Type I errors. This is why we present a third method, namely the studentized permutation test, which allows us to make inferences about the net benefit, Δ .

3.5.1 Unmatched net benefit

3.5.1.1 Randomization test

The most common approach to perform statistical inference for the net benefit is the randomization test (Buyse 2010). The null hypothesis of the test is H_0 : The joint distribution of the outcomes are the same VS H_1 : The joint distribution of the outcomes are not the same between the control and the treatment group. This is important because it mean that this test might reject the null hypothesis only because the distribution of the outcomes in both arm differ in variance while there is no difference in treatment effect. Here is how the test is run Gai (2021):

1. Estimate the net benefit $\hat{\Delta}$ on the original data
2. Simulate numerous times (S) a new vector of allocation based on the randomization used for the trial
3. For each simulation estimate $\hat{\Delta}_s$, $s \in \{1 \dots S\}$. The net benefit simulated will be different from zero only by chance under H_0 and therefore, the distribution of $\hat{\Delta}_s$ of all simulated samples is an approximation of the distribution of $\hat{\Delta}$ under H_0 .
4. The p-value can then be computed easily. For example, for a two-sided test, let s_2 be the number of $\hat{\Delta}_s$ such that $|\hat{\Delta}_s| \geq |\hat{\Delta}|$ $s \in \{1, \dots, S\}$ then the p-value is $\frac{s_2}{S}$

It is worth to note that this re-sampling method is computer intensive and that in order to have a precise estimation of the variance to the second digit at level of 5% 730'000 permutation samples are needed (Buyse 2010).

3.5.1.2 Asymptotic test

When the sample size is large enough, one can avoid the resampling method because the net benefit statistic ($\hat{\Delta}$) belongs to the class of U-statistics. Therefore, one can rely on asymptotic theory, we have that $\hat{\Delta}$ follows asymptotically a normal distribution. In the particular case of no treatment effect (under H_0), we get that $\Delta = 0$ and a two sided p-value is given by (Deltuvaite-Thomas and Burzykowski 2022):

$$p_{val} = 2 \cdot \Phi\left(\frac{-|\hat{\Delta}|}{\sqrt{\mathbb{V}(\hat{\Delta})}}\right) \quad (3.14)$$

Where Φ is the cumulative distribution of the standard normal distribution and an estimator of the variance is defined as follow :

$$\widehat{\mathbb{V}(\hat{\Delta})} = \frac{\sum_{i=1}^N (\sum_{j=1}^N W_{ij})^2}{(n_1 \cdot n_2)(n_1 + n_2)(n_1 + n_2 - 1)} \quad (3.15)$$

3.5.1.3 Studentized permutation test in paired data

When our data come from a setting such as measure at baseline and at end of the treatment period for each individuals, directly applying standard randomization or asymptotic tests to test if the treatment effect $\Delta = 0$ is inappropriate. This is because the data is paired in nature and considering that we have two independent group is a wrong assumption which violates the independence assumption that underpins asymptotic methods. Randomization tests also suffer in this scenario because it test the equality in the distribution of the outcomes between the two group which might differ only in term of variability. Moreover, they disregard the inherent dependence structure and because of the paired nature of the data we should resample at the individual level to conserve the correlation structure of the data, leading to substantial power loss and potentially too low Type I error. Section 5 empirically demonstrates these limitations of what we will call the "naive" approach.

The appropriate statistical tool to do inference on net benefit (Δ) for such data with finite sample size was proposed by Konietzschke and Pauly (2012) in the case of a single outcome for the probabilistic index that we have introduce in section 3.2. If we can do inference for the PI, we can use those result for the net benefit. This method focuses on testing the null hypothesis $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$. While Munzel (1999) developed an asymptotic test, its performance suffers in finite sample sizes. The Type I error can be inflated (for negative correlation) or deflated (for positive correlation) depending on the degree of correlation within the data.

We will first present Munzel's test with common rank-based notation, followed by how it connects to GPC methods in order to be able to find the variance estimator in the case of multiple outcome. Finally, we will introduce the studentized procedure by Konietzschke and Pauly (2012) which is valid for finite sample sizes. In the derivation presented below the normalized versions of the distribution functions are used $F_k(y) = \frac{1}{2}(F_k^+(y) + F_k^-(y))$ $k = 0, 1$ where $F_k^+(y) = P(Y \leq y)$ and $F_k^-(y) = P(Y < y)$. This is required in order to have the integration by parts rule $\int F_0 dF_1 = 1 - \int F_1 dF_0$ fulfilled when both samples are interchanged. On top of that, we will assume that each patient goes under both treatments we will therefore call N the number of patients as in the case of matched tests. To estimate Θ let's replace the distribution function by their empirical estimate

$$\hat{F}_k(y) = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} (\mathbf{1}\{Y_{ik} \leq y\} + \mathbf{1}\{Y_{ik} < y\}) \quad k = 0, 1. \quad (3.16)$$

Then an estimator for the PI is given by :

$$\hat{\Theta} = \int \hat{F}_0 d\hat{F}_1 = \frac{1}{N} \sum_{i=1}^N \hat{F}_0(Y_{i1}) = \frac{1}{N} (\bar{R}_{1.} - \bar{R}_{0.}) + \frac{1}{2}. \quad (3.17)$$

Here, $\bar{R}_{k.} = \frac{1}{N} \sum_{i=1}^N R_{ik}$ $k = 0, 1$ where R_{ik} is the rank of Y_{ik} among all $2N$ observations.

Munzel (1999) demonstrated that the rank statistic $T_n = N^{1/2}(\hat{\Theta} - \Theta)$ asymptotically follows a normal distribution with mean 0 and variance $\sigma_{\hat{\Theta}}^2 = \mathbb{V}(F_1(Y_{11}) - F_0(Y_{10}))$. To obtain a consistent estimator of the variance, define

$$Z_i = \frac{1}{N}(R_{i,1} - R_{i,1}^{(1)} - R_{i,0} + R_{i,0}^{(0)}) \quad (3.18)$$

where $R_{i,k}^{(k)}$ represents the internal rank of patient i within its group $k \in 0, 1$. The consistent variance estimator is then given by:

$$\hat{\sigma}_{\hat{\Theta}}^2 = \frac{1}{N-1} \sum_{i=1}^N (Z_i - \bar{Z})^2. \quad (3.19)$$

Under the null hypothesis $H_0 : \Theta = 1/2$, the distribution of the statistic

$$M_n = \frac{N^{1/2}(\hat{\Theta} - 1/2)}{\hat{\sigma}_{\hat{\Theta}}} \quad (3.20)$$

is asymptotically normal. For small sample sizes, it is advisable to approximate the distribution of M_n with a t-distribution with $2N-1$ degrees of freedom. However, Konietzschke and Pauly (2012) demonstrated that the accuracy of this approximation depends on the amount of inter-individuals correlation present in the data.

For proper inference in finite samples, irrespective of the correlation, the studentized test proposed by Konietzschke and Pauly (2012) should be used. The procedure is as follows:

- Compute the Munzel statistic: M_n as given by Equation (3.20).
- For $s \in \{1, \dots, S\}$, let $\tau_1, \dots, \tau_N$ be independent and identically distributed permutations of $\{(0, 1), (1, 0)\}$, and let $Y'_i = (Y_{i,\tau_i(1)}, Y_{i,\tau_i(2)})'$ denote the permuted pairs for patient $i = 1, \dots, N$. This involves randomly switching the group label for each patient and not doing random permutation between the two groups.
- For each s , compute the values of M_n^s and store them in a vector.
- The two-sided p-value is calculated as:

$$p_{\text{val}} = \min(2p_1, 2 - 2p_1)$$

$$\text{where } p_1 = \frac{1}{S} \sum_{s=1}^S \mathbb{1}\{M_n \leq M_n^s\}.$$

To ensure the validity of the procedure, the number of permutations should be at least 10,000. In cases where the variance estimators are zero, replace $\hat{\sigma}_{\hat{\Theta}}^2$ with $1/N$.

We will now make the link between the rank based approach and the pairwise comparison method. This will allow us to use the same inferential tool in the case of multivariate outcome, the presentation of this section is based on the work of Brunner (2024).

First, let's recall that an estimator of PI is given by $\hat{\Theta} = \frac{1}{2N}(\bar{R}_1 - \bar{R}_0) + \frac{1}{2}$ (see Equation 3.17).

The overall ranks R_{ik} can be split in two parts, one coming from comparing Y_{i0} and Y_{i1} with all observations within their group and one with all observations of the other group.

$$R_{i0} = \sum_{j=1}^N [\mathbb{1}\{Y_{j1} < Y_{i0}\} + \frac{1}{2}\mathbb{1}\{Y_{j1} = Y_{i0}\}] + \frac{1}{2} + \sum_{i=1}^N [\mathbb{1}\{Y_{s0} < Y_{i0}\} + \frac{1}{2}\mathbb{1}\{Y_{s0} = Y_{i0}\}]. \quad (3.21)$$

The first sum is the placements (R_{i0}^*) which represent the rank of Y_{i0} within the sample post treatment. The second term (including the $1/2$) is the internal ranks of Y_{i0} (R_{i0}^0). In a similar way we can define R_{i1}^* and R_{i1}^1 . Therefore, the overall ranks R_{ik} is simply the sum of the placements R_{ik}^* and the internal ranks R_{ik}^k .

Now if we introduce another comparisons function $c(x, y)$ which takes the following values :

$$c(x, y) = \begin{cases} 1, & y > x \\ 1/2, & y = x \\ 0, & y < x \end{cases} \quad (3.22)$$

It can be shown that $\hat{\Theta} = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N U_{ij}$, where $\mathbf{U}$ is a matrix of pairwise comparisons between groups, with element $u_{ij} = c(Y_{i0}, Y_{j1})$. The observations from the treatment group are placed in column, while the observations from control group are in row. To make it clear, the element $u(1, 1)$ represent the comparison between the patient 1 when he is under treatment vs. control. The element $u(1, 2)$ is the comparison between the patient 1 and patient 2 when patient 1 is under control while patient 2 is under treatment.

From this score matrix we can retrieve the placement $R_{i1}^* = R_{i,1} - R_{i,1}^{(1)}$ by taking the column sums $R_{i1}^* = u(+, j)$. By observing that $c(x, y) = 1 - c(y, x)$, the row sums $u(i, +)$ is equal to :

$$u(i, +) = \sum_{j=1}^N c(Y_{j0}, Y_{i1}) = \sum_{j=1}^N [1 - c(Y_{i1}, Y_{j0})] = N - R_{i0}^*. \quad (3.23)$$

Finally, the estimator $\hat{\Theta}$ is simply the mean of the matrix

$$\hat{\Theta} = u(+, +)/N^2. \quad (3.24)$$

We can retrieve the previous variance estimator (see Equation 3.19) by noticing that

$$Z_i = \frac{1}{N} (u(+, j) - (N - u(i, +))). \quad (3.25)$$

The main advantage of using pairwise comparisons is that it allows for a straightforward extension to the case where we use a threshold or in the case of multivariate outcome.

3.5.2 Matched net benefit

The matched net benefit (Γ) is a difference between two probabilities, namely the difference between the probability that a pair is a win ($\pi_w = P(Y_{i1} - Y_{i0} > \tau)$) minus the probability that a pair is a loss ($\pi_l = P(Y_{i1} - Y_{i0} < \tau)$), where we assume that larger is better. Γ is therefore given by:

$$\Gamma = \pi_w - \pi_l. \quad (3.26)$$

An estimator of these probabilities is given by $\hat{\pi}_w = \frac{1}{N} N_w$ and $\hat{\pi}_l = \frac{1}{N} N_l$, where N_w represents the number of wins and N_l the number of losses. The vector $(N_w, N_l, N_t)'$, where N_t represents the number of ties, follows a multinomial distribution of parameters $(N, \pi_w, \pi_l, 1 - \pi_w - \pi_l)$. The variance of the estimator is given by (Verbeek 2022):

$$\begin{aligned}
\sigma_{\hat{\Gamma}}^2 &= \mathbb{V} \left(\frac{1}{N} (N_w - N_l) \right) \\
&= \frac{1}{N^2} (N\pi_w(1 - \pi_w) + N\pi_l(1 - \pi_l) - 2\text{Cov}(N_w, N_l)) \\
&= \frac{1}{N^2} (N\pi_w(1 - \pi_w) + N\pi_l(1 - \pi_l) + 2N\pi_w\pi_l) \\
&= \frac{1}{N} (\pi_w + \pi_l - (\pi_w - \pi_l)^2).
\end{aligned} \tag{3.27}$$

Under the null hypothesis $H_0 : \Gamma = 0$, we can estimate the variance of the estimator by:

$$\hat{\sigma}_{\hat{\Gamma}}^2 = \frac{1}{N} (\hat{\pi}_w + \hat{\pi}_l) = \frac{1}{N^2} (N_w + N_l). \tag{3.28}$$

Matsouaka (2022) propose the following test statistic to test $H_0 : \Gamma = 0$:

$$Z = \frac{\hat{\pi}_w - \hat{\pi}_l}{\sqrt{\hat{\sigma}_{\hat{\Gamma}}^2}} = \frac{\sum_{i=1}^N W_i}{\sqrt{\sum_{i=1}^N |W_i|}}, \tag{3.29}$$

where Z follows a standard normal distribution. The type I error probability of this test statistic is demonstrated to be close to the desired significance level α . A major disadvantage of this test statistic is that it completely disregards the number of ties. For example the test would lead to the same p-value in the following two settings:

- 10 wins, 2 losses, and 0 ties.
- 10 wins, 2 losses, and 20 ties.

However, it seems evident that the first setting is a more favorable result.

In this section, we reviewed the main measures of treatment effect within the GPC framework. We then discussed how to estimate the matched and unmatched net benefit. We explored various methods for inference about the net benefit in the unmatched case and will demonstrate in Section 5 that only the studentized permutation test allows for proper inference while controlling the Type I error rate when that data is paired in nature. For the matched method, the situation is simpler, as the test proposed by Matsouaka (2022) remains valid regardless of the level of correlation between the baseline and end of treatment measurements.

4. Linear mixed model

4.1 Introduction

This thesis aims to evaluate the applicability of the GPC methodology to rare disease trials and assess the relative power of the matched versus (vs.) unmatched GPC approaches. As mathematical expressions for power are not easily derived, we will employ simulations to evaluate these performance metrics. Linear mixed models will serve as tools to generate data that mimic that of a rare disease trial. Furthermore, we will use these models to analyze data from the sirolimus trial, highlighting their limitations (see Section 6.3.2). It is common in clinical trial that the outcomes of interest are longitudinal, meaning they are measured repeatedly. To mimic the evolution of such results over time and take into consideration the likelihood of correlation between repeated observations on the same subject, one may employ a linear mixed model. For a longitudinal continuous outcome, a linear mixed model essentially assumes that each member of the population has a common trajectory and that over time, each member of the population deviates from this common trajectory to create a subject-specific profile that is exclusive to him. This section will examine only the random effects model. The literature on linear mixed model is large but it's not the objective of this work to provide a throughout treatment of that subject. This section is based on Legrand (2021).

4.2 Linear mixed effects model

To formally introduce the representation of longitudinal data, we let Y_{ij} denote the response of subject i , $i \in \{1, \dots, N\}$ at time t_{ij} , $j \in \{1, \dots, T_i\}$. The simplest model is a linear model with an intercept and a linear effect over time. However, different subjects tend to have different intercepts and slopes. Therefore, to take into account this heterogeneity between subjects we introduce random effects for the intercept and slope of each subject. The equation for the model is then given by :

$$Y_{ij} = \alpha_0 + \beta_{0i} + (\alpha_1 + \beta_{1i})t_{ij} + \epsilon_{ij}. \quad (4.1)$$

Where ϵ_{ij} represents the errors terms and $\beta_i = (\beta_{0i}, \beta_{1i})'$ is a vector of random effects. The assumption that β_i is independent of the ϵ_{ij} and that $\beta_i \sim \mathcal{N}(0, \Sigma)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ is frequently made. It is evident that the fixed parameter α_0 represents the average intercept where α_1 is the average slope. For each individual i , β_{0i} represents the random deviation in the intercept, and β_{1i} the random variation in the slope. The correlation between two measurement made on the same patient is induced by the random effect β . We can show that between two measurement points on the same patient the correlation is given by :

$$\rho(Y_{ij}, Y_{ij'}) = \frac{\sigma_{\beta_0}^2 + (t_{ij} + t_{ij'})\sigma_{\beta_0, \beta_1}}{\sqrt{\sigma_{\beta_0}^2 + t_{ij}^2\sigma_{\beta_1}^2 + 2t_{ij}\sigma_{\beta_0, \beta_1} + \sigma_{\epsilon}^2} \sqrt{\sigma_{\beta_0}^2 + t_{ij'}^2\sigma_{\beta_1}^2 + 2t_{ij'}\sigma_{\beta_0, \beta_1} + \sigma_{\epsilon}^2}} \quad (j \neq j'). \quad (4.2)$$

From this expression, we can see that even if the two random effects are independent, there will still be some correlation between the two measurements due to the random effect β_0 . However, measurements taken from two distinct patients remain independent.

It is possible to expand this model to incorporate more fixed and random factors. The generic statement of the linear mixed models is provided by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad (4.3)$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ represent the design matrices for the fixed and random effects, respectively. $\beta_i \sim \mathcal{N}(0, \Sigma)$ and $\epsilon_i \sim \mathbf{N}(0, R_i)$ with the common assumption that β_i is independent of ϵ_i . It's an usual assumption as well to treat the error terms as independent and homoskedastic leading to a matrix $R_i = \sigma_{\epsilon}^2 I_{T_i}$ with I_{T_i} the $T_i \times T_i$ identity matrix. However, alternative forms of R_i matrices, such as auto-regressive models, Toeplitz models, or unstructured models where the covariance matrix is left completely unspecified, might be taken into consideration if the random effects included in the data do not fully reflect the correlation in the data. These models are referred as covariance pattern models when there are no random effects included and the R_i matrix fully captures the correlation. Typically, the variables in the model are represented by $p + 1$, columns in the design matrix X , whereas the first column is a column of 1 representing the intercept and p is the number of covariates.

The $p + 1$ component of the coefficient vector $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)'$ where α_0 is the average intercept while α_l $l \in \{1, \dots, p\}$ represents the average change in the outcome Y_i when the corresponding covariates X_l increases by one unit, while the other covariates are kept constant. The design matrix $\mathbf{Z}$ of the random effects do not include an intercept. Random effects can be added to a subset of the regression parameters to capture how the i^{th} subject deviates from the average population parameters, with the same meaning as in the straightforward prior example.

One important point to remember is that the β_i 's are realization of a random variable coming from a normal distribution in our case. Consequently they rather refer to predictions. The parameters in Σ are the real parameters of interest linked to these random effects. We see that the number of parameters does not increase with the sample size which is a considerable advantage of the random effects over fixed effects model.

One of the main benefits of mixed models is that they can handle any imbalance in the data, which can arise from a variety of reasons, such as different numbers of measurements for different individuals or different time intervals for different individuals. This means that mixed models are not limited to balanced designs, where measurements are made for each individual at the same time points. Naturally, this reflects a more realistic scenario in practice. Additionally, the connection between many measurements taken on the same person can be considered using mixed-effects models. Indeed, because the measurements taken for the i^{th} person have the same random effects β_i , as demonstrated by the basic model, the model accounts for a possible correlation between them.

5. Simulation matched vs. unmatched GPC

5.1 Introduction

The objective of this section is to compare the performances of the matched and unmatched GPC approaches in terms of power in a setting where baseline and end of treatment values are available for each individual. As deriving closed-form mathematical expressions for the power of each test is not straightforward, we opted to investigate this through computer simulation. We utilized linear mixed models to generate data, aiming to create data generating processes (DGPs) that closely resemble real-world data, in our case specifically mimicking the data from the sirolimus trial. We investigate how each variance parameter affect the two net benefit (Γ and Δ), their estimation and our ability to make correct inference about it. The main goal is to be able to give recommendation on whether one should used the matched GPC approach or the unmatched one. We start with a single outcome namely the Visual Analog Scale (VAS) which is a subjective measure used to assess pain intensity. It consists of a 10-centimeter horizontal line, with endpoints representing extremes of pain (e.g., "no pain" to "worst imaginable pain"). Patients mark the line to indicate their perceived pain level, and the score is measured in centimeters from the left endpoint, providing a quantifiable measure of pain severity. We then run another simulation but by including this time a second endpoint, namely the number of crisis per months.

5.2 Single outcome

5.2.1 Sampling distribution

The DGP was the following, we set the mean VAS score to 5 and we draw the random effect (intercept and slope) from a bivariate normal of mean $\mathbf{0}$ and covariance matrix Σ . We set the average reduction in VAS score from baseline to the end to -1 and we add some random noise from a Gaussian distribution. As the VAS score is bounded between 0 and 10, we set values higher than 10 to 10 and values lower than 0 to 0. As the reported data are in a discrete scale we then round the data toward the nearest integer. The outcome for patient i is given by the following equation :

$$Y_{ij} = \mu + \beta_{0i} + (\lambda + \beta_{1i}) \cdot t_j + \epsilon_{ij}, \quad (5.1)$$

where :

- Y_{ij} : is the VAS score of patient $i \in \{1, \dots, N\}$ under treatment $j \in \{0, 1\}$
- μ : is the global mean at baseline (5 in the simulations)

- λ : is the slope, treatment effect (-1 in the simulations)
- $\beta_{0_i} \beta_{1_i}$: are the random effects for slope and intercept of patient i .

$$(\beta_0, \beta_1)' \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

$$\Sigma = \begin{pmatrix} \sigma_{\beta_0}^2 & \sigma_{\beta_0, \beta_1} \\ \sigma_{\beta_0, \beta_1} & \sigma_{\beta_1}^2 \end{pmatrix}$$

- ϵ_{ij} : is the random error following a normal $\mathcal{N}(0, \sigma_\epsilon^2)$
- $t_j = I(j = 1)$: is equal 0 at baseline and 1 after the treatment period.

For our simulations we study the impact of the following parameters :

- Sample size low (N=10) VS high (N=100)
- Parameter of variance of the intercept : low (1) VS high (5)
- Parameter of variance of the slope : low (1) VS high (5)
- Correlation intercept/slope : none (0) VS high (0.7)
- Variance of the error term : low (0.01) VS high (1)

The choice of those parameters was kind of arbitrary and could have been chosen differently. Nevertheless, we justify the setting of the treatment effect to -1 because in the preliminary trial they consider a small decrease in pain to be between -1 to -3 (Seront et al. 2023) and that from the available data we estimate that decrease to be in the order of 1.24. We choose to investigate the case when there is almost no patient available, case when N=10, and when there is substantial amount of patients sample size of N=100, because in the sirolimus trial there are 132 patients available. The Figure 5.1 present an illustration of one of the dataset generated for three different settings, all with the same variance of the random error, set to 1. On the top left the data are generated with all parameters of variance set to 1 and no correlation between the random effect of the intercept and the random effect of the slope. In the top right, the parameters of variance are set to 5 and no correlation between the random effect of the intercept and the random effect of the slope. On the bottom left the parameter of variance are set to 5 and the correlation between the random effect of the intercept and the random effect of the slope is set to 0.7. Finally on the bottom right, we display the true data from the sirolimus trial. We observe on the top right that the slope goes approximately in the same direction while on the top left the behaviour is more chaotic. By naked eyes it's hard to see a clear distinction between the DGP on the top right and the one on the bottom left. However, the variance of the VAS score at the end is 3 for the DGP of the dataset on the top right, while it's 11 for DGP on the top left, 19 for the DGP of the dataset on the bottom left and it's estimated to 7.37 for the data from the sirolimus trial. With this modeling of the VAS score, we can show that the correlation between the VAS at baseline and the VAS at the end is given by Equation (4.2) with $t_{ij} + t_{ij'} = 1$:

$$\rho(Y_{i0}, Y_{i1}) = \frac{\sigma_{\beta_0}^2 + \sigma_{\beta_0, \beta_1}}{\sqrt{\sigma_{\beta_0}^2 + \sigma_\epsilon^2} \sqrt{\sigma_{\beta_0}^2 + \sigma_{\beta_1}^2 + 2\sigma_{\beta_0, \beta_1} + \sigma_\epsilon^2}}. \quad (5.2)$$

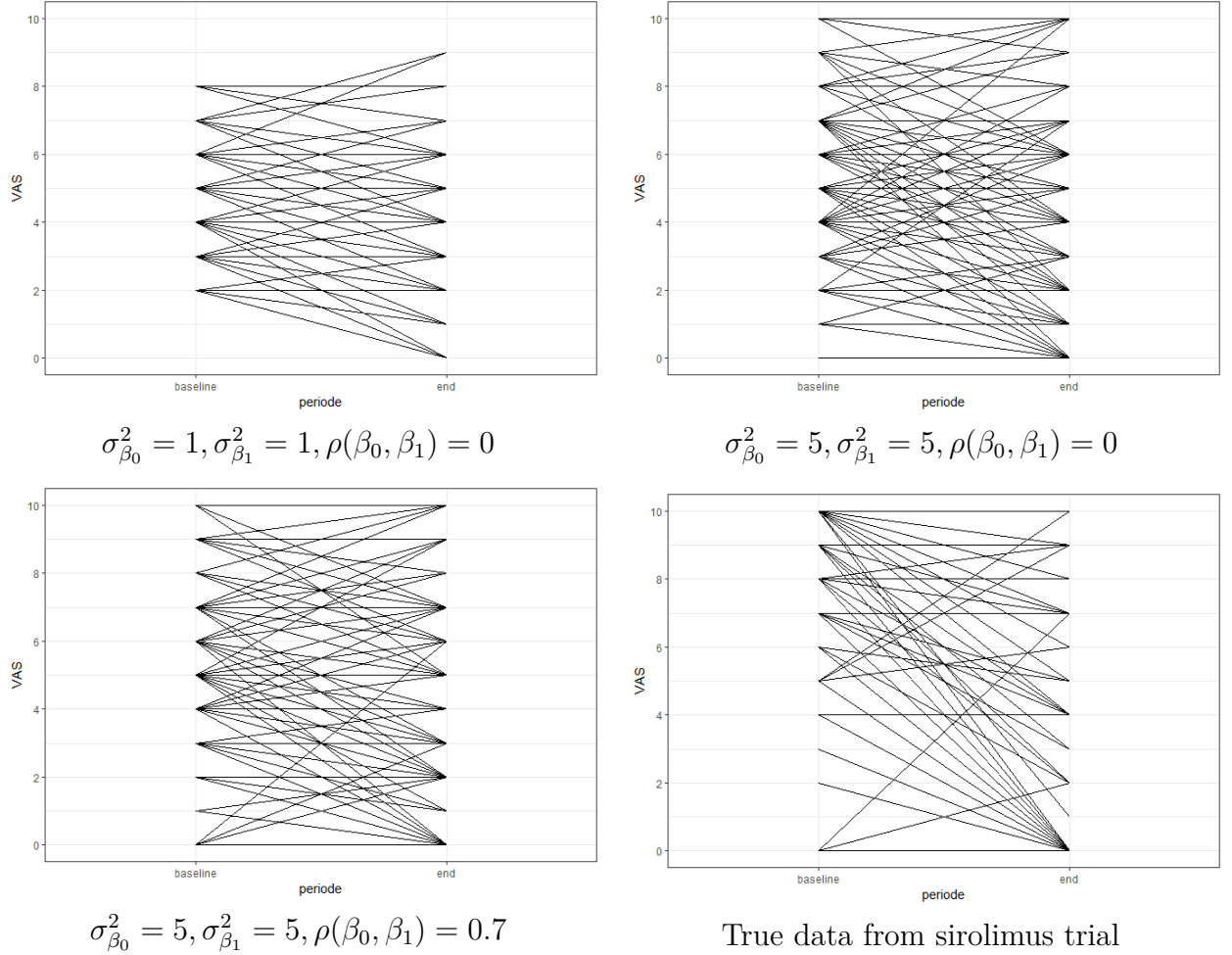


Figure 5.1: Illustration of one dataset generated for the simulations for different DGP and the data from the sirolimus trial each with 132 patients. On the top left one with low variability and no correlation between the random effect, on the top right one with high variability and no correlation between the random effect while on the bottom left one with high variability and high correlation between the random effect. Finally, on the bottom right the data collected during the sirolimus trial

We design a 2^5 full factorial design to study the effect of each parameter on the sampling distribution, we run 32 Mont-Carlo simulations, for each setting we generated 500 datasets where we compute each time the net benefit for the matched GPC and the unmatched GPC.

5.2.2 Theoretical net benefit

To gain more insight, it may be beneficial to first compute the theoretical net benefit for our various simulations. Typically, in simulations, the true values of the parameters of interest are known. However, since the net benefit is a difference between two probabilities (see Section 3.2), it isn't immediately evident and needs to be calculated given our DGPs. This section aims to derive the theoretical net benefit for both the unmatched GPC (Δ) and the matched version (Γ). In our setting, this is feasible because we utilize the normal distribution, which simplifies the calculations.

5.2.2.1 Unmatched net benefit

One can show that given our simulation model (see equation 5.1)

$$\Delta = 1 - 2 \cdot \mathbb{P}(Z < 0), \quad Z \sim \mathcal{N}(\lambda, 2\sigma_{\beta_0}^2 + 2\sigma_{\beta_0, \beta_1} + \sigma_{\beta_1}^2 + 2\sigma_{\epsilon}^2). \quad (5.3)$$

Proof. Let $\Delta = \mathbb{P}(Y_{i1} > Y_{j0}) - \mathbb{P}(Y_{i1} < Y_{j0})$

$$\begin{aligned} \Delta &= \mathbb{P}(Y_{i1} > Y_{j0}) - \mathbb{P}(Y_{i1} < Y_{j0}) \\ &= \mathbb{P}(\mu + \beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} > \mu + \beta_{0j} + \epsilon_{jk'}) - \mathbb{P}(\mu + \beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} < \mu + \beta_{0j} + \epsilon_{jk'}) \\ &= \mathbb{P}(\beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} - \beta_{0j} - \epsilon_{jk'} > 0) - \mathbb{P}(\beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} - \beta_{0j} - \epsilon_{jk'} < 0) \\ \Omega &= \begin{pmatrix} \beta_{0i} \\ \beta_{1i} \\ \beta_{0j} \\ \epsilon_{ik} \\ \epsilon_{jk'} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\beta_0}^2 & \sigma_{\beta_0, \beta_1} & 0 & 0 & 0 \\ \sigma_{\beta_0, \beta_1} & \sigma_{\beta_1}^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{\beta_0}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\epsilon}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\epsilon}^2 \end{pmatrix} \right) \end{aligned}$$

$$\text{Let } s = (1 \quad 1 \quad -1 \quad 1 \quad -1)'$$

Then $W = s' \Omega \sim \mathcal{N}(0, 2\sigma_{\beta_0}^2 + 2\sigma_{\beta_0, \beta_1} + \sigma_{\beta_1}^2 + 2\sigma_{\epsilon}^2)$ linear combination of the normal distribution

By letting $Z = W + \lambda$

We have $\Delta = \mathbb{P}(Z > 0) - \mathbb{P}(Z < 0) = 1 - 2\mathbb{P}(Z < 0)$ Symmetry of the normal distribution

□

5.2.2.2 Matched net benefit

One can show that given our simulation model (see equation 5.1)

$$\Gamma = 1 - 2 \cdot \mathbb{P}(Z < 0), \quad Z \sim \mathcal{N}(\lambda, \sigma_{\beta_1}^2 + 2\sigma_{\epsilon}^2). \quad (5.4)$$

Proof. Let $\Gamma = \mathbb{P}(Y_{i1} > Y_{i0}) - \mathbb{P}(Y_{i0} < Y_{i1})$

$$\begin{aligned}
\Gamma &= \mathbb{P}(Y_{i1} > Y_{i0}) - \mathbb{P}(Y_{i0} < Y_{i1}) \\
&= \mathbb{P}(\mu + \beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} > \mu + \beta_{0i} + \epsilon_{ik'}) - \mathbb{P}(\mu + \beta_{0i} + \lambda + \beta_{1i} + \epsilon_{ik} < \mu + \beta_{0i} + \epsilon_{ik'}) \\
&= \mathbb{P}(\lambda + \beta_{1i} + \epsilon_{ik} - \epsilon_{ik'} > 0) - \mathbb{P}(\lambda + \beta_{1i} + \epsilon_{ik} - \epsilon_{ik'} < 0) \\
\Omega &= \begin{pmatrix} \beta_{1i} \\ \epsilon_{ik} \\ \epsilon_{ik'} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\beta_1}^2 & 0 & 0 \\ 0 & \sigma_{\epsilon}^2 & 0 \\ 0 & 0 & \sigma_{\epsilon}^2 \end{pmatrix} \right) \\
\text{Let } s &= (1 \quad 1 \quad -1)' \\
\text{Then } W &= s' \Omega \sim \mathcal{N}(0, \sigma_{\beta_1}^2 + 2 \cdot \sigma_{\epsilon}^2) \text{ linear combination of the normal distribution} \\
\text{By letting } Z &= W + \lambda \\
\text{We have } \Gamma &= \mathbb{P}(Z > 0) - \mathbb{P}(Z < 0) = 1 - 2\mathbb{P}(Z < 0) \text{ symmetry of the normal distribution}
\end{aligned}$$

□

We can make a couple of concluding remarks concerning the previous 2 results. First, we see that the net benefit for the unmatched version depends on $\sigma_{\beta_0}^2$ multiplied by a factor of 2 and on the covariance between β_0 and β_1 . So we expect that increasing the variability of the intercept should have more impact on reducing the net benefit than increasing the variance of the slope. And that increasing the correlation between the two parameter should reduce the net benefit. Secondly, we observe that the net benefit of the matched GPC does not depend on $\sigma_{\beta_0}^2$ and neither on the correlation between β_0 and β_1 .

5.2.3 Power

To study the power of each methods we run 1000 monte carlo simulations, we use the same data generating process as mentioned above. We test the following hypothesis :

In the case of unmatched GPC :

$$H_0 : \Delta = 0 \text{ vs. } H_1 : \Delta \neq 0$$

In the case of matched GPC :

$$H_0 : \Gamma = 0 \text{ vs. } H_1 : \Gamma \neq 0$$

5.2.3.1 The naive approach

Before exploring the power, we need to ensure that the inferential technique we are using properly controls the Type I error. In this subsection, we will demonstrate that the first two techniques, which we refer to as naive approaches, are not appropriate for making inferences about Δ .

The naive approach 1

To evaluate Type I error rates, we generated data as equation 5.1 but we set λ equal to 0 as λ is the treatment effect. We conducted a simulation study encompassing both unmatched and matched GPC settings. For unmatched GPC, we employed studentized resampling methods for inference when the sample size (n) was less than 30. Resampling methods were used for n between 30 and 50, and for n greater than or equal to 50, we relied on asymptotic results. The R package “Buyse” was used for these analyses. In the case of matched GPC, we

adopted the methodology proposed by Matsouaka (2022) (see Section 3.5.2). The simulations explored a range of sample sizes (n) from 10 to 100, with increments of 5 at each step.

It's important to note that during our study of the unmatched GPC, we have a violation of the independence assumption. In fact, N out of N^2 paired are not independent because they come from the same individual. With those simulations, we wanted to check whether we can ignore this issue if the sample size was large enough, assuming that those pairs would be diluted and not be a concern.

The results of the type I error simulations are presented in Figure 5.2. We can see that the results for the matched method proposed by Matsouaka (2022) is around the right level of 5%. 5% is the significance level that we used, meaning that for each dataset generated we have applied the matched and unmatched test and we reject the null hypothesis of no treatment effect if the p-value obtained was below 5%. However, we see that the unmatched approach has a control of type I error which is too strict. Especially when the variance of the slope random effect is set to 5 while keeping the variance of slope parameter to 1. We think that this phenomenon occurs because the data is paired in nature. Using the naive method does not work for the unmatched method, probably because the independence assumption, on which the test procedure relies, is not met, as we mention in Section 3.5.

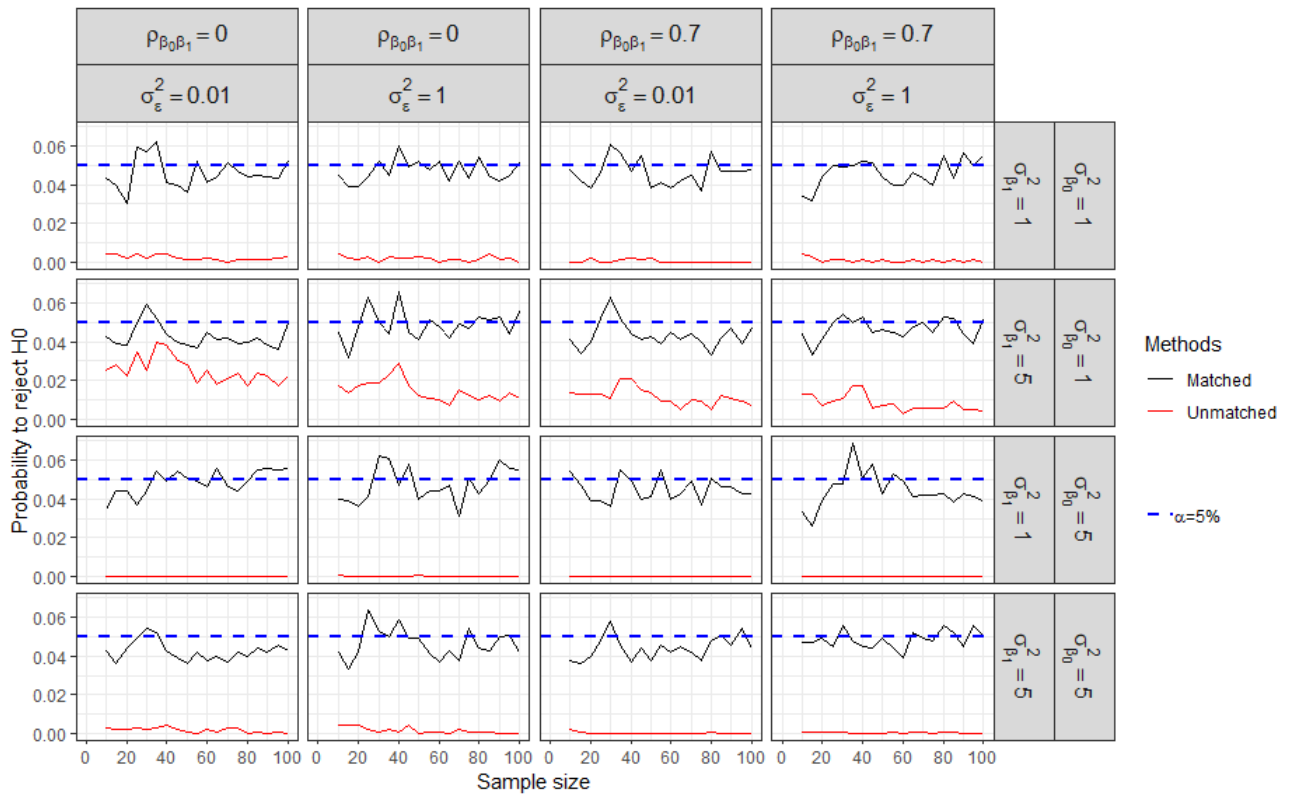


Figure 5.2: Monitoring of type I error for the matched and unmatched GPC under various values of the variance parameters

The naive approach 2

Another idea might be to keep the same test method but to permute the data at the individual levels in order to keep the correlation structure of the data. Nevertheless, given our simulation setup this will not work as well as shown in Figure 5.3. The randomization test tests for the equality between the two distributions, the introduction of the random

effect β_1 creates extra variability at the second time period and therefore the test rejects too often. On top of that, with this methodology we are testing the equality in distribution between the control and treatment group, but we want to test $\Delta = 0$. This is why we needed to use studentized Munzel test for which the error of type I is controlled as shown in the next section.

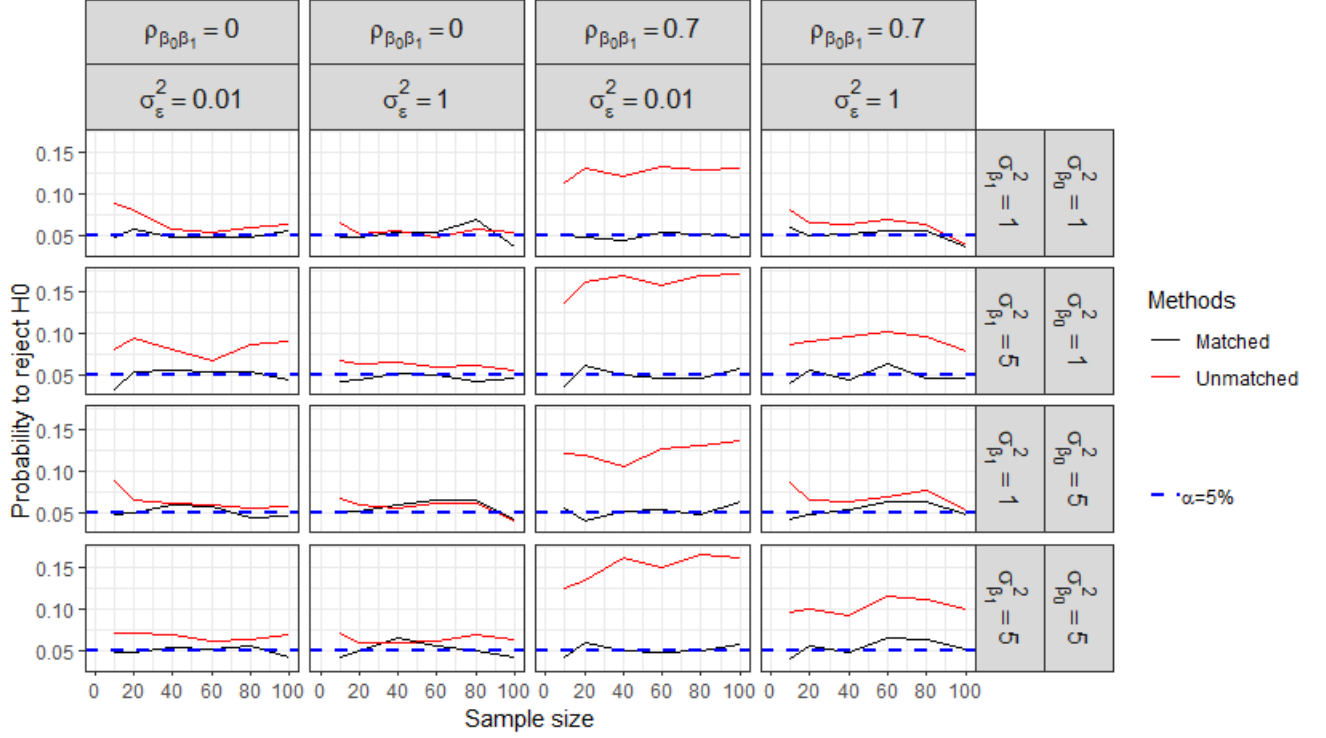


Figure 5.3: Monitoring of type I error for the matched and unmatched GPC under various values of the variance parameters when permuting only at the individual level

5.2.3.2 Results for the comparison of the matched vs unmatched GPC

The Figures 5.4 and 5.5 present our results, comparing the sampling distribution of the net benefit estimator for both methods. When the sample size is set to 10, we observe consistently higher variability in the estimate for the matched methods compared to the unmatched one when the variability of the random effects for the intercept and slope are set to 1. However, when we increase the variance of the random effect of the slope, both methods exhibit similar variability around their mean estimates. Similarly, maintaining the variance of the random effect for the slope at 1 but increasing the variability of the random effect for the intercept results in comparable variability around their respective estimates, similarly to the case when both variances are set to 1. Notably, there is a shift in the mean estimation of the net benefit for the unmatched method from around 0.4 to 0.2 in this scenario. While the matched method is between 0.6 to 0.5 across the various scenario.

When both random effects' variances are set to 5, we observe similar variability for both methods. It's worth noting that the estimation for the net benefit in the unmatched method decreases as the variance of the random effect increases, while the estimation for the matched GPC remains relatively stable at around 0.5, irrespective of the scenario. The impact of noise does not seem to significantly affect either method, and the effect of correlation only serves to reinforce the preceding conclusions. As we show in our theoretical derivation increasing

the variance of the intercept has more effect on the value of the net benefit (towards lower value) for the unmatched approach than increasing the variance of the slope. However, we see that uncertainty around the estimation is higher when the variance parameter of the slope increase rather than the intercept.

The effect of the sample size further reinforces our previous conclusions, indicating smaller variability for parameters when variances are low and a shift in the net benefit estimation from 0.45 toward 0.35 for the unmatched method and from 0.6 toward 0.5 for the matched approach.

Table 5.1 shows the comparison of the theoretical vs empirical net benefit of our simulation (N=100). We can see that our theoretical result aligns with the simulations.

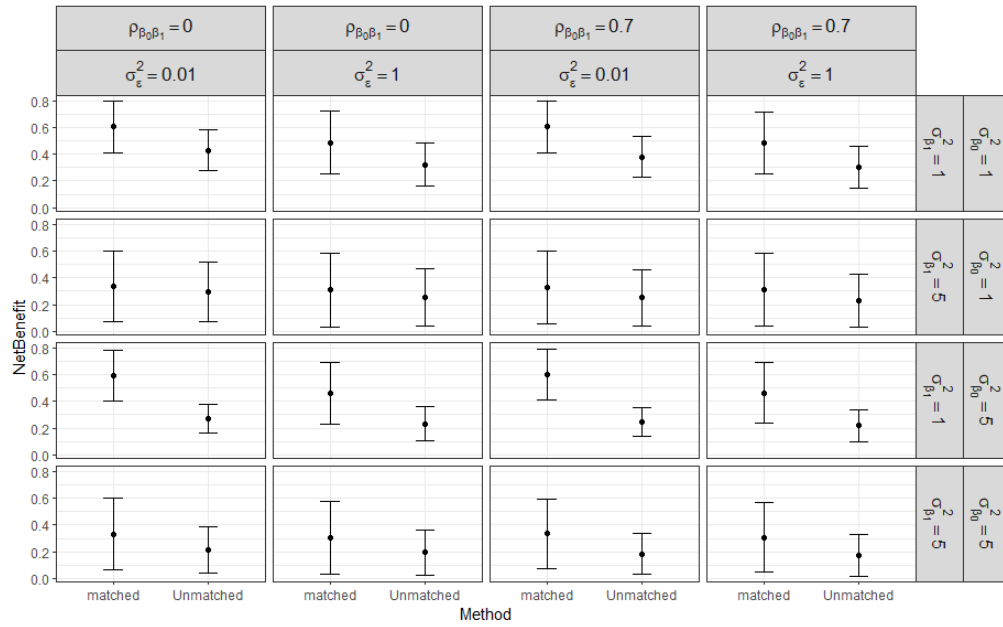


Figure 5.4: Comparison of the net benefit for the matched GPC vs. unmatched GPC for different values of the variance parameters when the sample size is set to 10, the bar represents the mean +/- one standard deviation

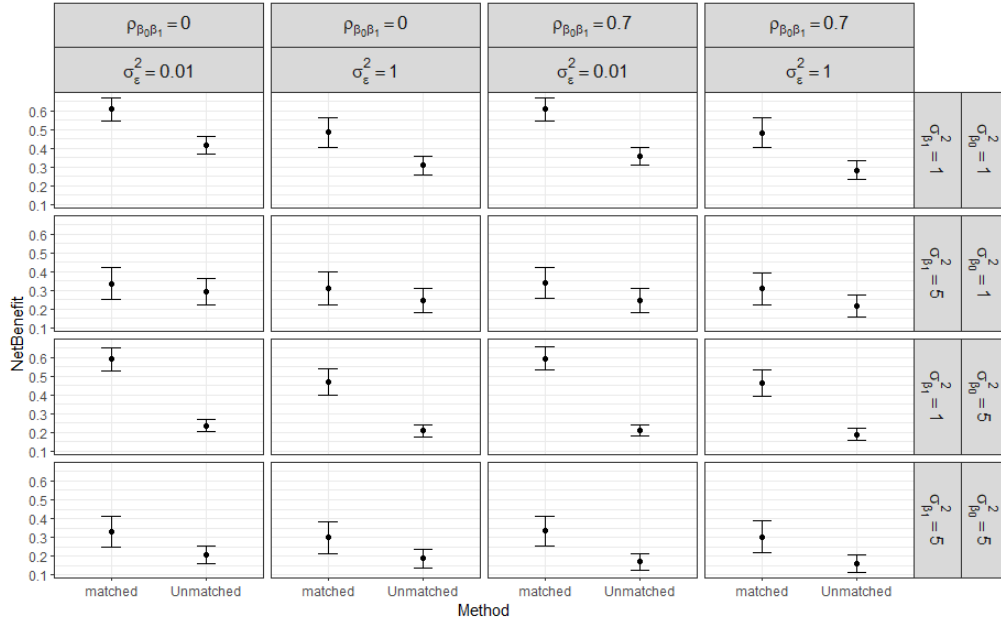


Figure 5.5: Comparison of the net benefit for the matched GPC vs. unmatched GPC for different values of the variance parameters when the sample size is set to 100, the bar represents the mean $\pm$ one standard deviation

Table 5.1: Empirical vs. theoretical derivation of the the net benefit for the matched and unmatched GPC for various configuration.

Parameters				Unmatched GPC		Matched GPC		Corr(Y_{i0}, Y_{i1})
$\sigma_{\beta_0}^2$	$\sigma_{\beta_1}^2$	ρ	σ_{ϵ}^2	Empirical Δ	Theoretical Δ	Empirical Γ	Theoretical Γ	
1	1	0	0.01	0.434	0.435	0.677	0.678	0.701
1	1	0	1	0.322	0.345	0.525	0.436	0.408
1	1	0.7	0.01	0.369	0.366	0.678	0.678	0.916
1	1	0.7	1	0.291	0.307	0.523	0.436	0.573
1	5	0	0.01	0.295	0.294	0.349	0.345	0.405
1	5	0	1	0.249	0.241	0.317	0.295	0.267
1	5	0.7	0.01	0.245	0.246	0.343	0.345	0.844
1	5	0.7	1	0.221	0.226	0.315	0.295	0.569
5	1	0	0.01	0.241	0.237	0.668	0.678	0.911
5	1	0	1	0.215	0.218	0.507	0.436	0.771
5	1	0.7	0.01	0.214	0.210	0.674	0.678	0.970
5	1	0.7	1	0.192	0.197	0.512	0.436	0.842
5	5	0	0.01	0.204	0.204	0.337	0.345	0.706
5	5	0	1	0.186	0.192	0.305	0.295	0.615
5	5	0.7	0.01	0.171	0.192	0.349	0.345	0.920
5	5	0.7	1	0.159	0.162	0.313	0.295	0.817

The results of the Type I error and power simulations are presented in Figure 5.6 and Figure 5.7. We observe that the Type I error is properly controlled for both methods, irrespective of the correlation between the two endpoints. Consequently, we can compare the power of both methods.

Regarding power, we observe that both methods perform similarly regardless the amount of the correlation between the two measures. From the plot, we see that the power for both tests is primarily influenced by the variance of the error term, σ_ϵ^2 , and the variance of the slope parameter, $\sigma_{\beta_1}^2$. When these two parameters are set to their high values, the number of patients needed to achieve a power of 80% exceeds 50. However, when $\sigma_{\beta_1}^2$ and σ_ϵ^2 are set to their low values, only 20 patients are required to reach a power of 80%.

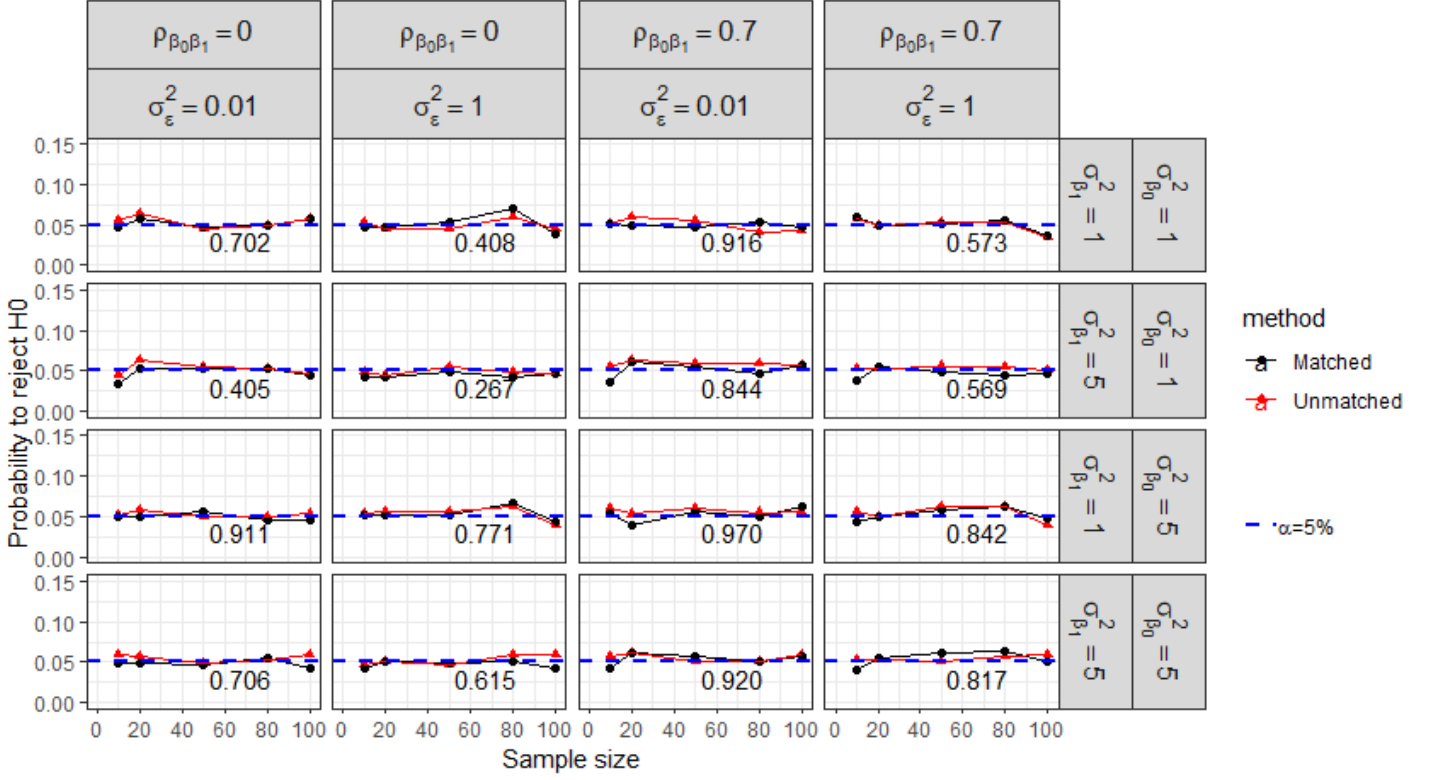


Figure 5.6: Type I error simulations of unmatched GPC vs. matched GPC inference for different values of the variance parameters when the true treatment effect (λ) is set to 0 for testing hypothesis of the kind $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ and $H_0 : \Gamma = 0$ vs. $H_1 : \Gamma \neq 0$. The number on the bottom left of each panel represents the correlation between the two the two measures (baseline/end of treatment period) induced by the random effect

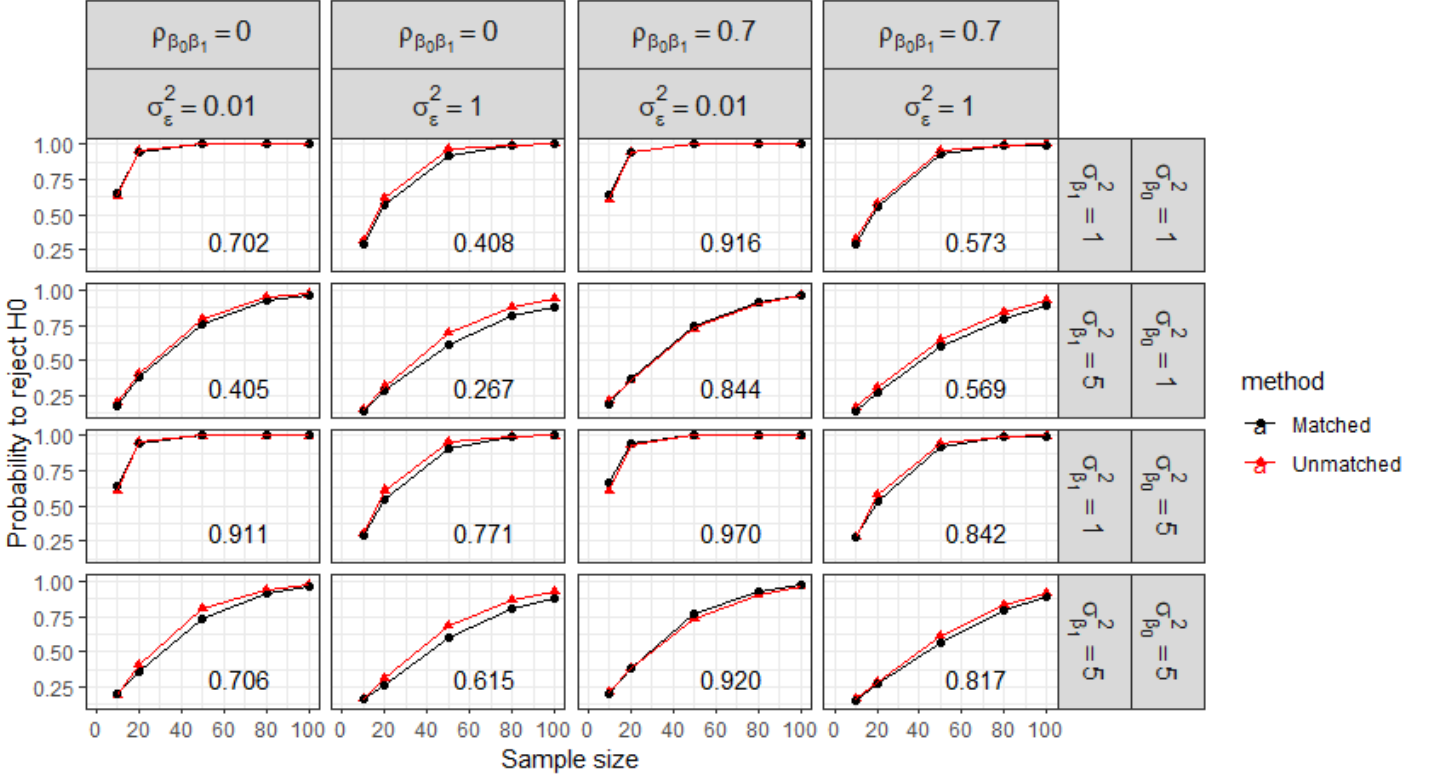


Figure 5.7: Power comparison of unmatched GPC vs. matched GPC for different values of the variance parameters when the true treatment effect (λ) is to -1 for testing hypothesis of the kind $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ and $H_0 : \Gamma = 0$ vs. $H_1 : \Gamma \neq 0$. The number on the bottom left of each panel represents the correlation between the two measures (baseline/end of treatment period) induced by the random effect

5.2.4 Extreme scenario

From the previous results we notice that the differences in power between the two methods were small. We saw that the setting corresponding to $\sigma_{\beta_0}^2 = 5, \sigma_{\beta_1}^2 = 1, \rho_{\beta_0, \beta_1} = 0$ and $\sigma_\epsilon^2 = 1$ was the one where the unmatched method was outperforming the most the matched method. On the other hand, we saw that the scenario with $\sigma_{\beta_0}^2 = 5, \sigma_{\beta_1}^2 = 5, \rho_{\beta_0, \beta_1} = 0.7$ and $\sigma_\epsilon^2 = 0.01$ was slightly in favor of the matched method. We decided to push those two scenarios to their extreme in order to see if they were a real gain in power for one method in comparison to the other. For the first scenario we set the variance parameter of β_0 near to zero in order to have the correlation between the two endpoints near to 0.

For the second extreme scenario, we investigate what could happen in case of a large correlation between the two measures (baseline/end of treatment period). For this we increase the parameter ρ_{β_0, β_1} from 0.7 to 0.99. In this case the correlation between the two measures was near 0.996. The value of the various parameters for the two scenarios are presented in Table 5.2.

Table 5.2: Value of the various parameters for the extreme case, the first scenario corresponds to the one where the two measures (baseline/end of treatment period) are not correlated at all, while the second setting correspond to a really high amount of correlation between the two measures (0.996).

Parameters	Extreme 1	Extreme 2
$\sigma_{\beta_0}^2$	0.0001	5
$\sigma_{\beta_1}^2$	1	5
ρ_{β_0, β_1}	0	0.99
σ_{ϵ}^2	1	0.01
ρ_{Y_0, Y_1}	0	0.996

Figure 5.8 shows the results for both scenarios. We observe that when there is no correlation between the measures, the unmatched approach outperforms the matched GPC. However, the improvement is relatively marginal. For example, with a sample size of 80 patients, we achieve a power of 90% with the unmatched GPC, compared to 85% with the matched GPC.

On the other hand, when the correlation is high, as in setting number two, we see that the matched method performs better than the unmatched one, although the improvement remains marginal. It is worth noting that for both methods, the power drops significantly (below 75%) when the sample size is below 60 patients for setting 1 and 50 patients for setting 2.

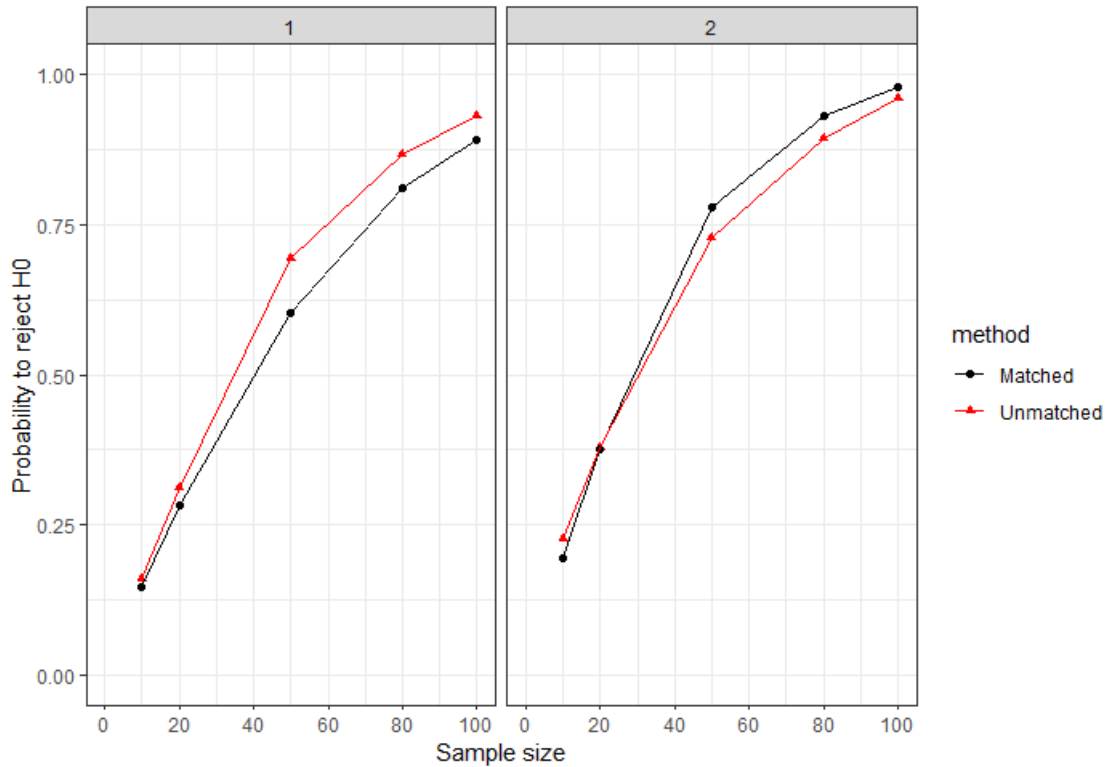


Figure 5.8: Power comparison of the matched GPC versus unmatched GPC for extreme settings is shown in Figure 5.8. On the left (1), the correlation between the endpoints is near 0, while on the right (2), we have a correlation between the endpoints of 0.996.

5.3 Two continuous outcomes

5.3.1 Simulation setup

The objective of this section is to extend the single outcome case by adding an extra outcome. So if we have tie in a pair we will look at the second outcome to see if the pair can be classified as favorable, neutral or a loss. Let Y_{ijk} be the value of outcome $k \in \{1, 2\}$ at time $j \in \{0, 1\}$, where 0 means at baseline and 1 at the end of the treatment period of patient $i \in \{1, 2, \dots, N\}$. Concerning the treatment effect λ_k we set the effect to -1 for both outcomes for one of the simulation and to -0.5 for the first outcome and -1 for the second set of simulations see appendix A. The data is generated from a linear mixed model as follow :

$$\begin{aligned} Y_{i01} &= \mu_1 + \beta_{0i} + \epsilon_{i01} \\ Y_{i11} &= \mu_1 + \beta_{0i} + \lambda_1 + \beta_{1i} + \epsilon_{i11} \\ Y_{i02} &= \mu_2 + \beta_{2i} + \epsilon_{i,0,2} \\ Y_{i12} &= \mu_2 + \beta_{2i} + \lambda_2 + \beta_{3i} + \epsilon_{i12} \\ \beta &\sim \mathcal{N}(\mathbf{0}, \Sigma) \\ \epsilon_{ijk} &\sim \mathcal{N}(0, \sigma_{\epsilon_k}^2). \end{aligned} \tag{5.5}$$

The correlation between the different measures on the same individual are induced by the random effects, we have that :

$$\begin{aligned} \rho(Y_{i,0,1}, Y_{i,1,1}) &= \frac{\sigma_{\beta_0}^2 + \sigma_{\beta_0, \beta_1}}{\sqrt{\sigma_{\beta_0}^2 + \sigma_{\epsilon_1}^2} \sqrt{\sigma_{\beta_0}^2 + \sigma_{\beta_1}^2 + 2\sigma_{\beta_0, \beta_1} + \sigma_{\epsilon_1}^2}} \\ \rho(Y_{i,0,2}, Y_{i,1,2}) &= \frac{\sigma_{\beta_2}^2 + \sigma_{\beta_2, \beta_3}}{\sqrt{\sigma_{\beta_2}^2 + \sigma_{\epsilon_2}^2} \sqrt{\sigma_{\beta_2}^2 + \sigma_{\beta_3}^2 + 2\sigma_{\beta_2, \beta_3} + \sigma_{\epsilon_2}^2}} \\ \rho(Y_{i,0,1}, Y_{i,0,2}) &= \frac{\sigma_{\beta_0, \beta_2}}{\sqrt{\sigma_{\beta_0}^2 + \sigma_{\epsilon_1}^2} \sqrt{\sigma_{\beta_2}^2 + \sigma_{\epsilon_2}^2}} \\ \rho(Y_{i,1,1}, Y_{i,1,2}) &= \frac{\sigma_{\beta_0, \beta_2} + \sigma_{\beta_0, \beta_3} + \sigma_{\beta_2, \beta_2} + \sigma_{\beta_1, \beta_3}}{\sqrt{\sigma_{\beta_0}^2 + \sigma_{\beta_1}^2 + 2\sigma_{\beta_0, \beta_1} + \sigma_{\epsilon_1}^2} \sqrt{\sigma_{\beta_2}^2 + \sigma_{\beta_3}^2 + 2\sigma_{\beta_2, \beta_3} + \sigma_{\epsilon_2}^2}}. \end{aligned} \tag{5.6}$$

We will use the previously identified scenarios to design the new simulations, as follows:

1. **Favoring the unmatched method for both outcomes:** We expect this scenario to favor both outcomes for the unmatched method. It corresponds to use the value of the parameters from the extreme scenario 1 (see previous Section 5.2.4).
2. **Favoring the Unmatched method for Outcome 1 and the Matched method for Outcome 2:** We expect this scenario to favor outcome 1 towards the Unmatched method and Outcome 2 towards the matched method. It corresponds to use the value of the parameters from extreme scenario 1 for outcome 1 and those from extreme scenario 2 for outcome 2 (see previous Section 5.2.4).
3. **Favoring the matched method for outcome 1 and the unmatched method for outcome 2:** We expect this scenario to favor outcome 1 towards the matched method and outcome 2 towards the Unmatched method. It corresponds to use the value of the parameters from extreme scenario 2 on outcome 1 and from extreme scenario 1 on outcome 2 (see previous Section 5.2.4).

4. **Favoring the matched method for both outcomes:** We expect this scenario to favor both outcomes in favor of the matched method. It corresponds to use the value of the parameters from extreme scenario 2 on both outcomes (see previous Section 5.2.4).

We still need to explain how we handle the correlation between the two outcomes. By choosing to design our simulation using the previous scenarios, we impose certain constraints on the matrix Σ , as it needs to be positive definite ($\Sigma \succ 0$). We have decided to take the following approach: set the correlation between Y_{i01} and Y_{i02} to either a low value of 0 or a high value of approximately 0.9.

Next, we determined the amount of correlation we wanted to induce between Y_{i11} and Y_{i12} , aiming to achieve as much correlation as possible, either positive or negative. We found the values of the variance parameters by solving the following convex optimization problem :

$$\begin{aligned} \min/\max f(\Sigma) &= \frac{\sigma_{02} + \sigma_{03} + \sigma_{12} + \sigma_{13}}{\sqrt{\sigma_0^2 + \sigma_1^2 + 2 \cdot \sigma_{01}} \sqrt{\sigma_2^2 + \sigma_3^2 + 2 \cdot \sigma_{23}}} \\ \text{such that } \Sigma &\succ 0 \\ \text{where } \Sigma &= \begin{bmatrix} \sigma_{\beta_0}^2 & \sigma_{\beta_0, \beta_1} & \sigma_{\beta_0, \beta_2} & \sigma_{\beta_0, \beta_3} \\ \sigma_{\beta_0, \beta_1} & \sigma_{\beta_1}^2 & \sigma_{\beta_1, \beta_2} & \sigma_{\beta_1, \beta_3} \\ \sigma_{\beta_0, \beta_2} & \sigma_{\beta_1, \beta_2} & \sigma_{\beta_2}^2 & \sigma_{\beta_2, \beta_3} \\ \sigma_{\beta_0, \beta_3} & \sigma_{\beta_1, \beta_3} & \sigma_{\beta_2, \beta_3} & \sigma_{\beta_3}^2 \end{bmatrix}. \end{aligned} \quad (5.7)$$

Note that given that we have already fixed some parameters, the only variables in the optimization problem are $\sigma_{\beta_0, \beta_3}$, $\sigma_{\beta_1, \beta_2}$, and $\sigma_{\beta_1, \beta_3}$. The full simulation plan with the values of the variance parameters is presented in Appendix A.

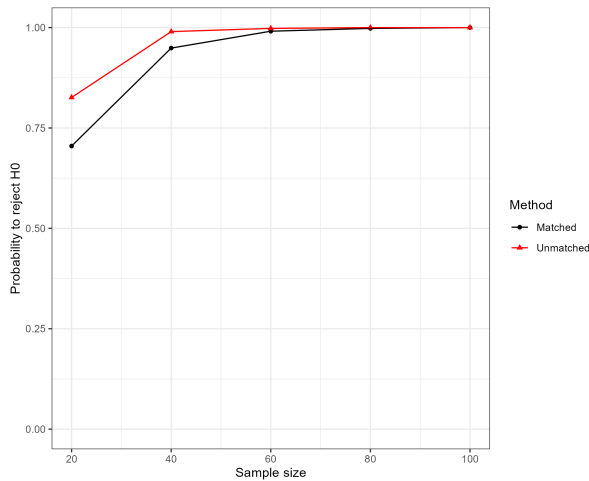
Concerning the inferential tools, we employed the test proposed by Matsouaka (2022) for the matched GPC. For the unmatched GPC, we utilized the studentized procedure proposed by Konietzke and Pauly (2012) when the sample size was below 80. When the sample size was 80 or 100, we applied the Munzel test (Munzel 1999).

5.3.2 Results for the comparison of the matched vs. unmatched GPC

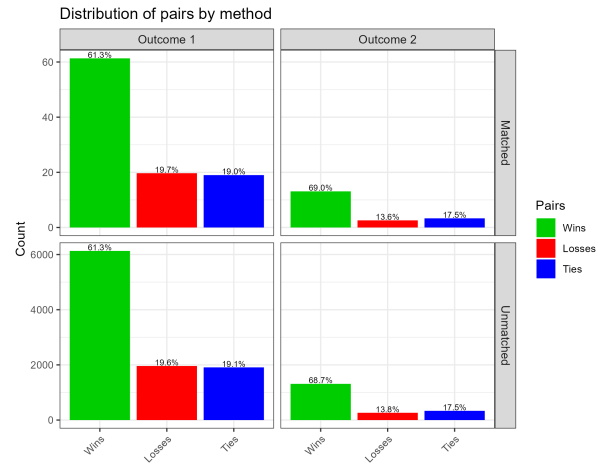
We begin by examining a scenario that we believe (based on our simulations of single outcome) must favor the unmatched method for both outcomes. This mean that we set the values of the parameters of variance to the one of extreme scenario 1 (see Section 5.2.4). We impose a negative correlation of -0.5 between outcome 1 and outcome 2 at the end of the treatment periods (See simulation plan in the Appendix 5.3.2). Figure 5.9 presents the results for this scenario. In the top section of the figure, the treatment effect is set to -1 for both outcomes, whereas in the bottom section, the treatment effect is set to -0.5 for outcome 1 and -1 for outcome 2. The left panel depicts the power for both methods, while the right panel illustrates the distribution of win pairs (green), loss pairs (red), and ties (blue) for both outcomes. It is noteworthy that, in contrast to the univariate case, the unmatched method demonstrates superior performance compared to the matched method, even with a smaller sample size. When the treatment effect is set to -1 for both outcomes, we observe a high power even with a low sample size, where the power for the unmatched method is around 80%, compared to 70% for the matched GPC. The distribution of pairs is consistent across methods: for the first outcome, we see 60% win pairs, 20% loss pairs, and 20% ties. Among the 20% tied pairs in the first outcome, 70% are wins in the second outcome, 13% are losses, and 17% remain ties.

Conversely, when the treatment effect for outcome 1 is reduced to -0.5 , there is a significant loss of power, especially with a low sample size. In this scenario, the power drops to approximately 40% for the unmatched method and 30% for the matched GPC when the sample size is set to 20. The pairs distribution remains the same across both methods: for outcome 1, 50% are win pairs, 29% are loss pairs, and 21% are ties. Within the ties from outcome 1, 65% are wins, 16% are losses, and about 19% are ties in outcome 2.

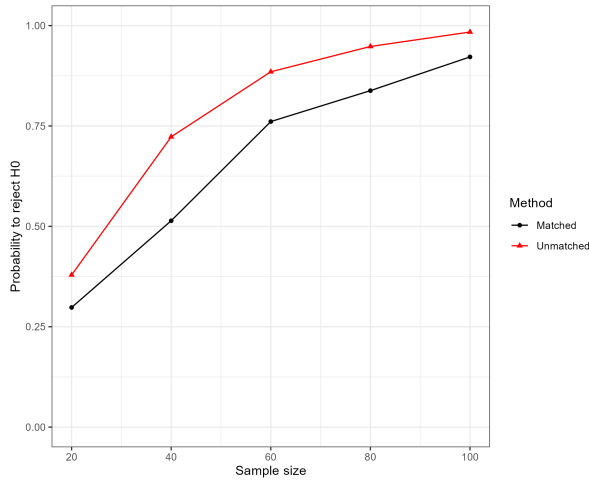
The reduction in the treatment effect for outcome 1 results in a higher proportion of unfavorable pairs in outcome 1 and does not significantly increase the number of tie pairs (only by 1%). For outcome 2, the pairs distribution remains almost identical, which is logical since the treatment effect for outcome 2 was not altered.



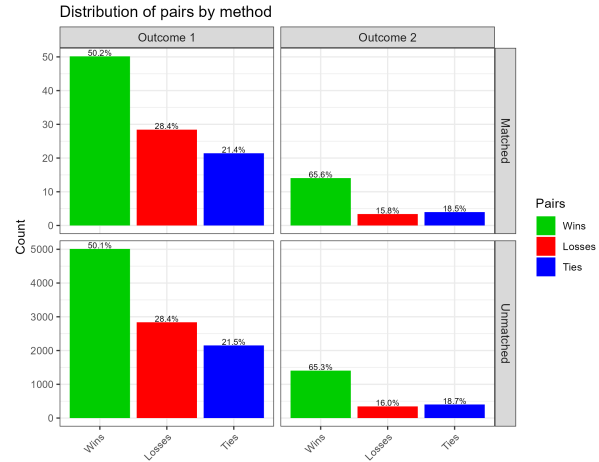
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes



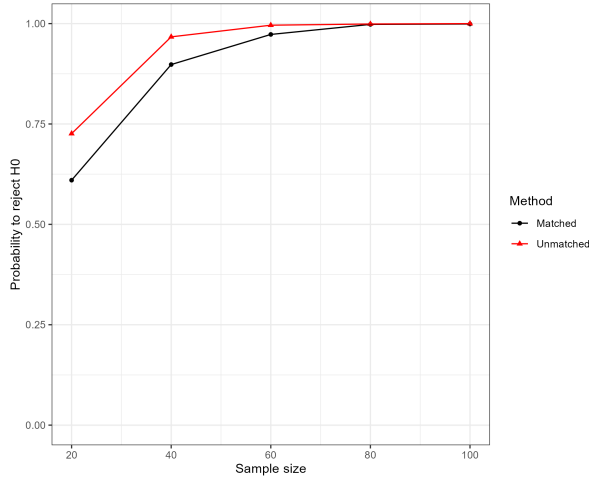
(c) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2



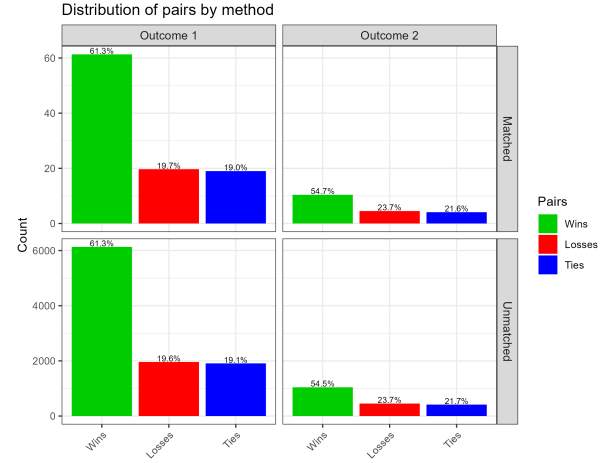
(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

Figure 5.9: Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on both outcomes, meaning that we used the values of the parameters from extreme scenario 1 (see Section 5.2.4) with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 1 in Appendix A).

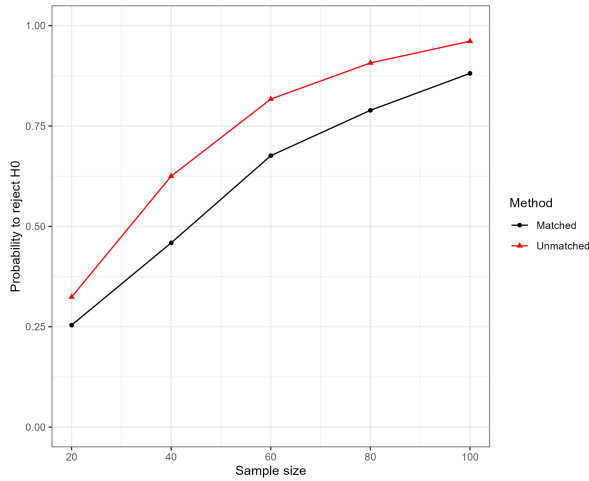
When we transition from a negative correlation between outcome 1 and outcome 2 at the end of the treatment period to a positive correlation, while keeping all other parameters constant, we observe a slight decrease in power, as illustrated in Figure 5.10, we see that the curves are shifted downward compared to Figure 5.9. This trend was consistent across all simulations. A detailed comparison of the pairs distributions between the two simulations reveals that the variation occurs only in the second outcome, as shown in Figure 5.11. Specifically, we see more wins when the correlation is negative.



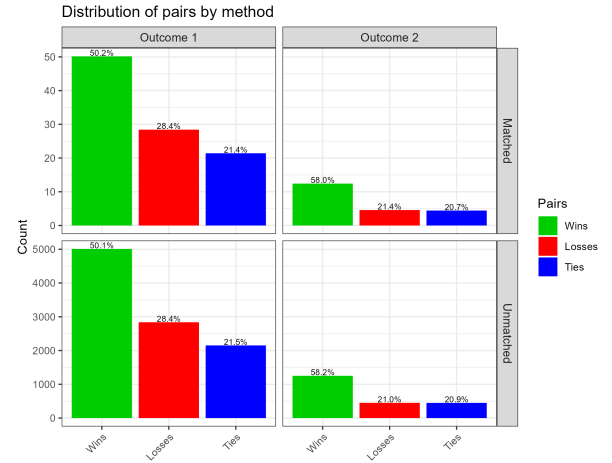
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes



(c) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2



(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

Figure 5.10: Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on both outcomes, meaning that we used the values of the parameters from extreme scenario 1 (see Section 5.2.4). With positive correlation between the two outcomes at the end of the treatment period (see simulation plan setup 2 in Appendix A).

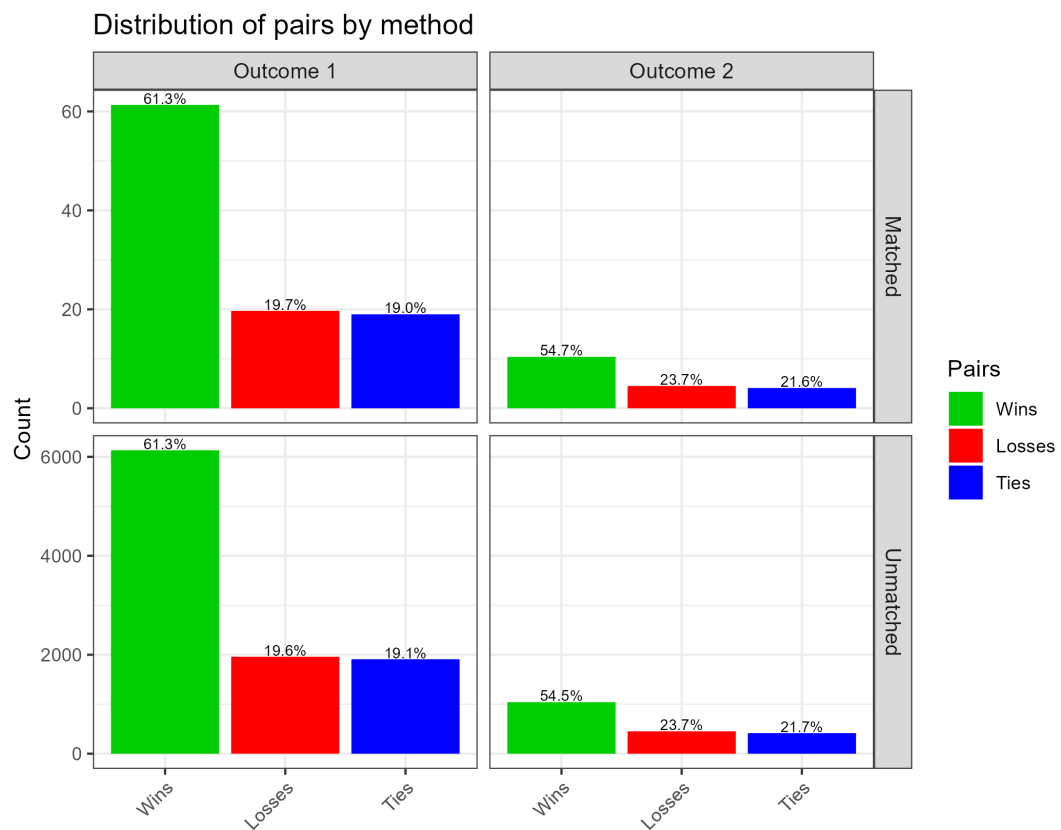
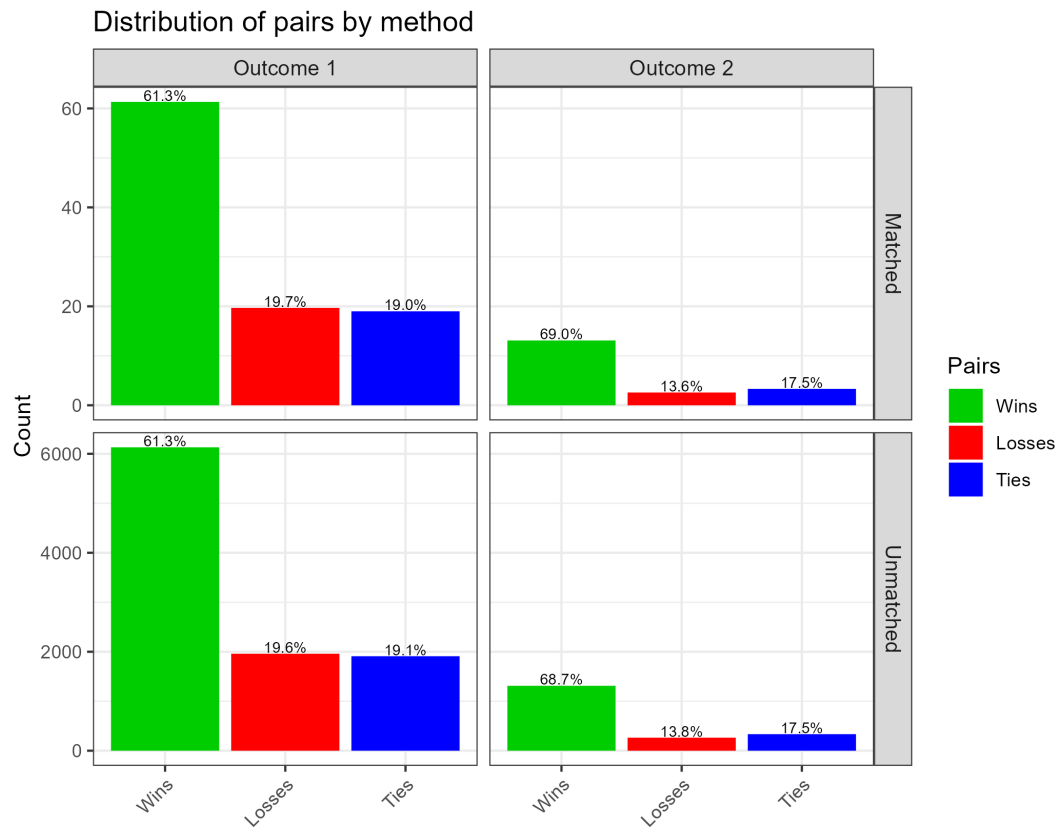


Figure 5.11: Both simulations have the same setup, namely the parameters form extreme scenario 1 (see Section 5.2.4) favoring the unmatched methods for both outcomes with a treatment effect of -1 on both outcomes but with the amount of correlation between outcome 1 and outcome 2 at the end of the treatment period set to -0.5 on top and to 0.5 on the bottom. (see simulation setup 1 and 2 in the Appendix A)

The intuition behind this observation is as follows: When the correlation between endpoints is negative, a tie occurs for a patient who didn't fare well on the first endpoint (scoring average or below average). Due to the negative correlation, we then move to the second endpoint, where these patients are expected, on average, to have a better outcome. This increases the likelihood of observing a 'win' for the treatment on the second endpoint, ultimately leading to increased power for the analysis. Conversely, when the correlation is positive, patients who tied on the first endpoint (due to average or below-average scores) are likely to continue having poor outcomes on the second endpoint as well. This reduces the chances of observing a 'win', thus decreasing the power of the analysis. In essence, the negative correlation creates a scenario where ties on the first endpoint are more likely to be resolved favorably on the second endpoint, contributing to the increased power observed in such cases.

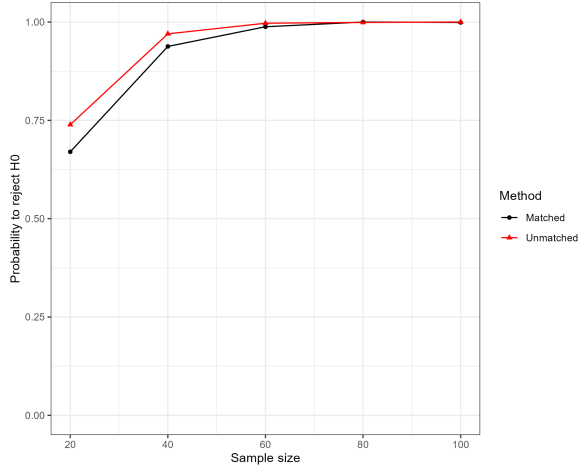
In simulation setups 3 and 4 (see Appendix A), we designed the conditions that must favor the unmatched methods (parameters values similar as those from extreme scenario 1 of Section 5.2.4) for the first outcome and the matched GPC for the second outcome (parameters values similar as those from extreme scenario 2 of Section 5.2.4). As a result, we observe a decrease in the power of the unmatched methods, bringing the power of the matched and unmatched methods closer together. This effect is even more pronounced when the treatment effect on the first outcome is reduced to -0.5 instead of -1, as shown in Figure 5.12.

The distribution of pairs for outcome 1 remains consistent with that observed with the simulations setup 1. However, for outcome 2, there is a significant difference in the distribution between the matched and unmatched methods. Specifically, the matched GPC method exhibits about 30% tie pairs, whereas the unmatched method shows approximately 13%. This difference can be attributed to the high correlation (0.99) for outcome 2 before and after the treatment period. Because the matched method compares pairs only at the individual level, it results in a higher number of ties due to the high correlation, which makes outcomes more similar within individuals. In contrast, the unmatched methods compare pairs both between and within individuals, but the correlation does not influence the pairing between individuals, leading to a lower proportion of tie pairs. This disparity highlights the nuanced differences in how the matched and unmatched methods handle highly correlated outcomes, ultimately affecting their respective power and effectiveness in different scenarios.

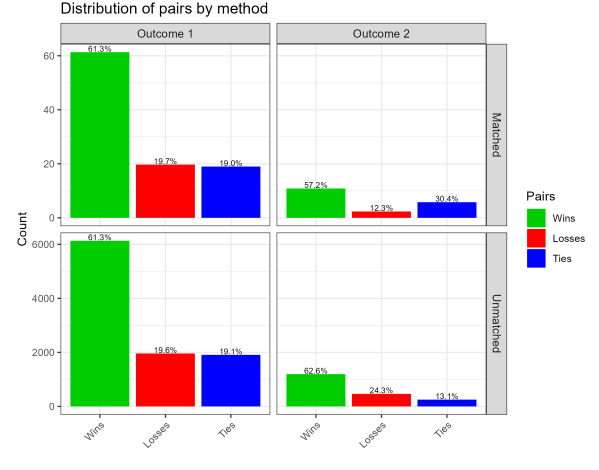
Next, we examine the scenario where the matched GPC method is supposed to be favor for outcome 1 by introducing a high correlation before and after the treatment period (parameters values from extreme scenario 2 of Section 5.2.4), while the unmatched GPC method is supposed to be favor for outcome 2 (parameters values from extreme scenario 1 of Section 5.2.4) see simulation plan setups 5 and 6 in Appendix A. By favoring the matched method on outcome 1, we observe a slight decrease in power, and the performance of the two methods begins to be similar, as illustrated in Figure 5.13.

The main difference in the distribution of pairs between the two methods lies in the percentages of losses and ties. For outcome 1, the matched method results in 48% win pairs, 30.6% loss pairs, and 21.4% ties, whereas the unmatched method shows 49.8% win pairs, 41.7% loss pairs, and 8.5% ties. However, on outcome 2 the distribution of the pairs is relatively similar between the two methods.

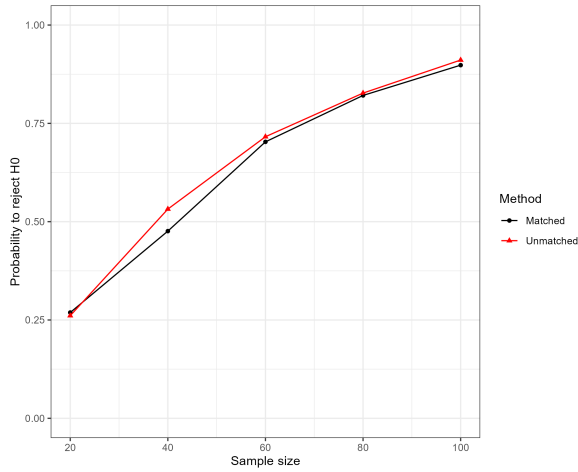
Following this, we examine the scenario where a high correlation is introduced before and after the treatment period for both outcomes, while keeping the two outcomes themselves uncorrelated, as outlined in simulation plan setups 7 and 8 (see Appendix A). We observe that for small sample sizes, namely 20 and 40, both methods exhibit similar power. However, as the sample size increases, the matched GPC method outperforms the unmatched method, as shown in Figure 5.14.



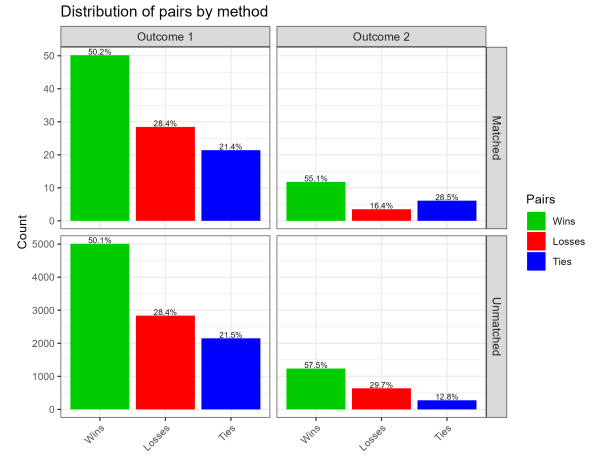
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes



(c) Power comparison of the matched GPC vs. Unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2

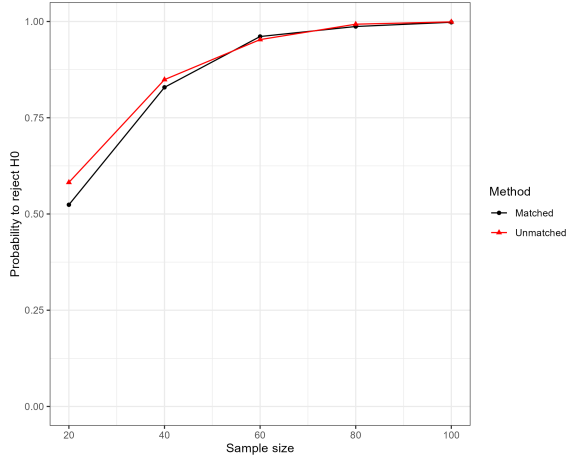


(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

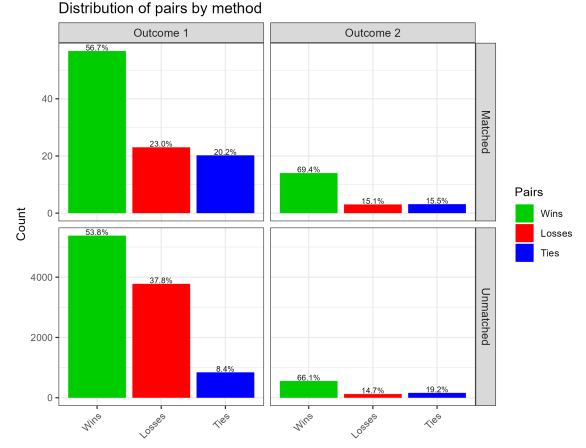
Figure 5.12: Comparison of the matched vs. unmatched GPC when we favor the unmatched methods on outcome 1 and the matched GPC on outcome 2 with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 3 in Appendix A)

It is important to note that, overall, there is a decrease in power, highlighting the trend that increasing the correlation before and after the treatment period tends to reduce the power of the methods. This suggests that higher intra-individual correlation may introduce complexities that make it more challenging for both matched and unmatched methods to detect treatment effects effectively, thereby diminishing their statistical power.

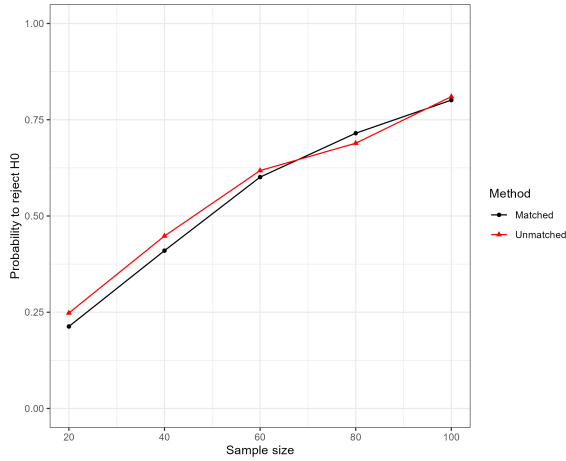
Finally, we examine a scenario similar to the previous one but with a high correlation between the two outcomes at baseline. Setting the correlation to 0.9 between the two outcomes at baseline imposes constraints on the achievable correlation at the end of the treatment period, which ranged from a minimum of 0.605 to a maximum of 0.726, as outlined in simulation plans 9 and 10 in Appendix A. Figure 5.15 illustrates the results, showing a



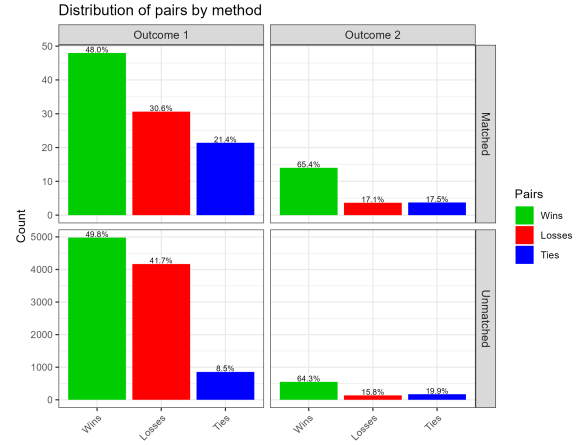
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes



(c) Power comparison of the matched GPC vs. Unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2



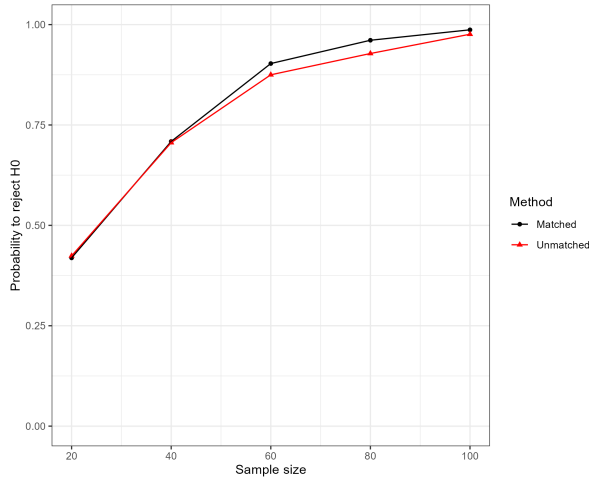
(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

Figure 5.13: Comparison of the matched vs. unmatched GPC when we try to favor the matched methods on outcome 1 with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 5 in Appendix A)

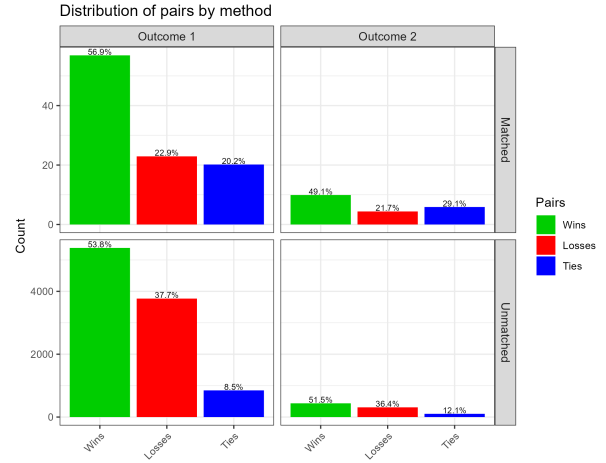
significant drop in power compared to Scenario 1, with the maximum power reaching only 55% at a sample size of 100.

Additionally, for small sample sizes, the matched method slightly underperform compared to the unmatched one; however, this trend reverses once the sample size reaches 60. The distribution of pairs between the two methods differs for both outcomes. While the percentage of win pairs is approximately the same for both methods, the matched method exhibits a higher proportion of tie pairs, 20% for outcome 1 and 30% for outcome 2. While the unmatched method has only 8% tie pairs for outcome 1 and 22% for outcome 2. This disparity underscores how baseline correlation between outcomes can significantly influence the effectiveness of matched and unmatched methods, particularly in terms of their ability to classify pairs as informative or tie.

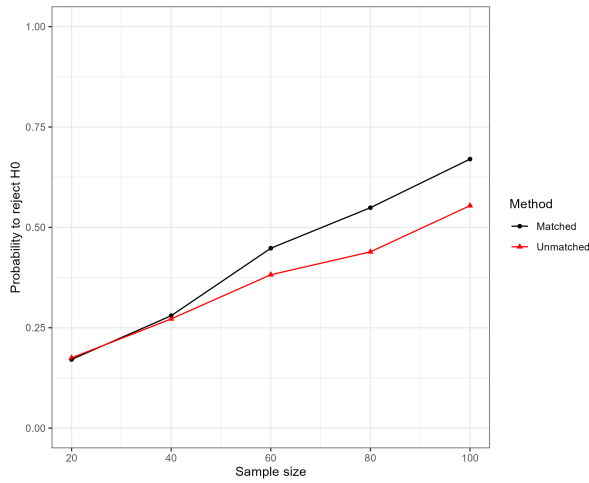
A closer examination of Figure 5.15 shows a noticeable drop for the unmatched method



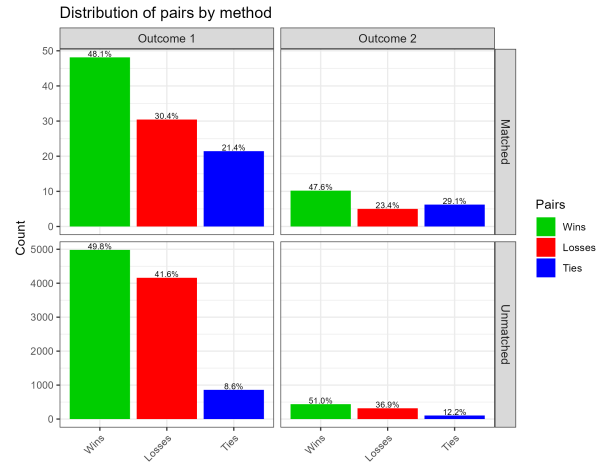
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes



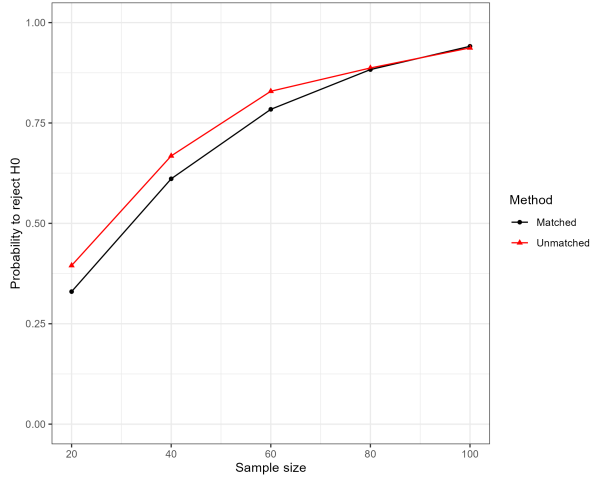
(c) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2



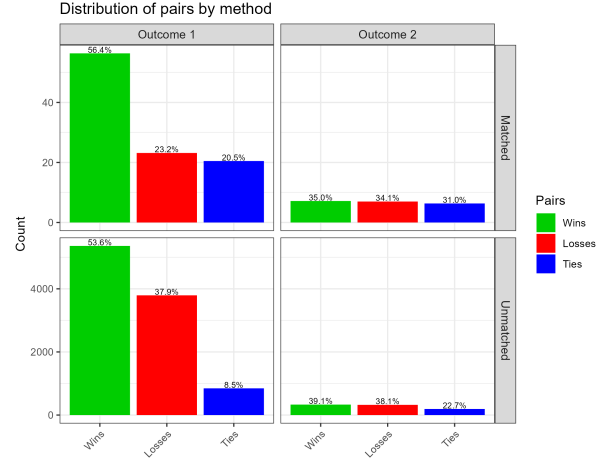
(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

Figure 5.14: Comparison of the matched vs. unmatched GPC when we favor the matched methods on both outcomes (using the values of parameters from extreme scenario 2 from Section 5.2.4) with negative correlation between the two outcomes at the end of the treatment period (see simulation plan setup 7 in Appendix A)

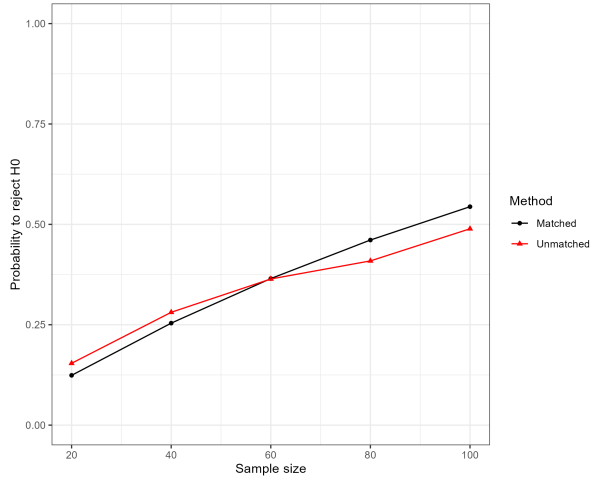
at a sample size of 80. This drop corresponds to the transition from non-asymptotic to asymptotic results. However, these differences should not be interpreted as significant because, as demonstrated in Figure 5.16, this drop would not occur if the studentized permutation test were applied across all sample sizes. While this choice may be subject to criticism, it was made to manage the computational burden, as running all simulations without relying on asymptotic results would have been excessively time-consuming.



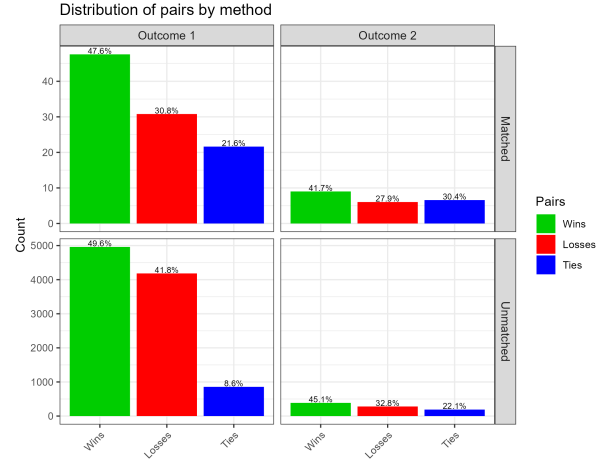
(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -1 on both outcomes



(b) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -1 on both outcomes

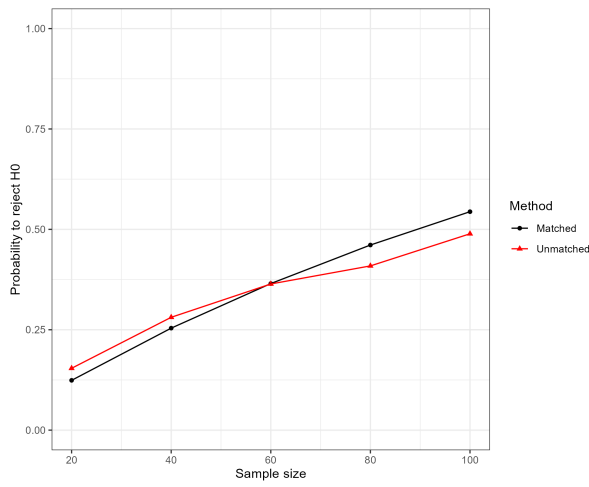


(c) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on outcome 1 and -1 on outcome 2

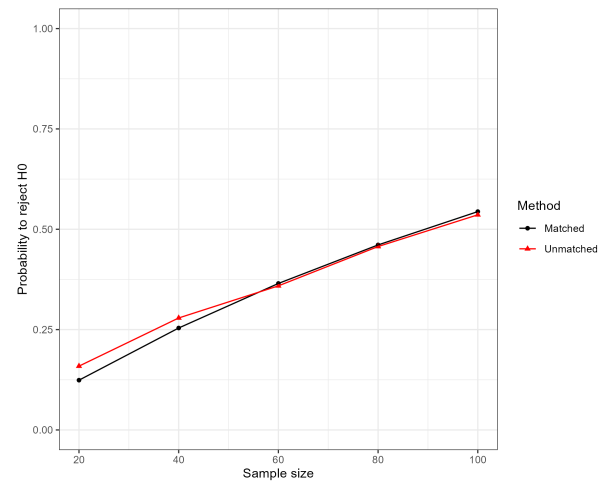


(d) Distribution of pairs for the matched and unmatched GPC with treatment effect set to -0.5 on outcome 1 and -1 on outcome 2

Figure 5.15: Comparison of the matched vs. unmatched GPC when the measures at baseline and at the end of treatment are highly correlated, using the values of parameters from extreme scenario 2 from Section 5.2.4 for both outcomes. The measures at baseline of outcome 1 and outcome 2 are highly correlated (see simulation plan setup 9 in Appendix A).



(a) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on the first outcomes and -1 on outcomes 2 with asymptotic inference for the unmatched method when the sample is 80 or above



(b) Power comparison of the matched GPC vs. unmatched GPC when the treatment effect is set to -0.5 on the first outcomes and -1 on outcomes 2 without asymptotic inference for the unmatched method

Figure 5.16: Comparison of the matched vs. unmatched GPC when the measures at baseline and at the end of treatment are highly correlated, using the values of parameters from extreme scenario 2 from Section 5.2.4 for both outcomes. The measures at baseline of outcome 1 and outcome 2 are highly correlated (see simulation plan setup 9 in Appendix A).

5.4 Discussion

Our simulations provide insight into how the patterns of correlation present in the data impact the power of GPC methods. In the univariate case, we demonstrate that when the measure at baseline and at the end of the treatment period are nearly independent, using the unmatched GPC method yields a slight gain in statistical power. This gain, however, is modest, with the maximum increase observed in a sample size of 60, where the unmatched GPC achieved around 70% power compared to 62% for the matched GPC. This modest gain in power is surprising as we have N^2 pairs for the unmatched and only N pairs for the matched method. But it is crucial to recognize that the two methods lead to different net benefit that are fundamentally different. The matched method net benefit (Γ) is much easier to interpret and communicate, as it measures the probability that an individual will experience a more favorable potential outcome under treatment than at baseline. In contrast, the unmatched net benefit (Δ) represents the expected improvement in terms of quantiles across the entire population (baseline and treatment) for an individual under treatment compared to baseline.

It is important to note that in practical scenarios, we would not typically observe two completely independent measurements on the same individual, implying some level of correlation. This correlation would diminish the power of the unmatched GPC resulting in similar performance between the two methods. Consequently, decision-makers might prefer to report a metric that is more easily interpretable, such as Γ .

Our findings in the bivariate case were consistent with the univariate case. When the two measures (baseline/end of treatment period) for both outcomes were nearly independent, the unmatched method outperformed the matched GPC to a greater extent than in the univariate scenario. Conversely, when there was significant correlation for the measures at baseline and at the end of the treatment period within each outcome, we observed a substantial drop in power and matched GPC seems more appropriate. Therefore, our first recommendation is that the more correlation one expects between baseline and end-of-treatment measurements, the larger the patient sample needed to achieve reasonable power. In our simulation, with extremely correlated outcomes, we observed that the power could be as low as 50% even with 100 patients when the treatment effect on the first outcome was set to -0.5 .

An interesting observation was the change in the distribution of pairs for each method with varying levels of correlation. With no correlation, both methods shared approximately the same distribution of pairs. However, as correlation increased, the distributions diverged, with the matched method exhibiting a higher number of tie pairs. This is logical since the matched method compares pairs at the individual level, leading to more ties due to high correlation.

6. Application to the sirolimus trial data

6.1 Introduction

Slow-flow vascular malformations (SFVMs) are a complex group of malformations that are caused by aberrant blood vessel development forming abnormal venous and lymphatic channels that become permanent. This group englobes Lymphatic malformation (LM), capillary malformations (CM) and venous malformations (VM), each with distinct clinical symptoms. An example of the different clinical symptoms is for example compressible, blue skin lesions, sometimes containing phleboliths (calcifications within the malformation) for patients with VMs. Patients with LMs typically present abnormal lymph-filled cysts, categorized as macrocystic (large cysts) or microcystic (small cysts) based on their size. LMs are typically present at birth and can affect any organ or tissue and are associated with complications that can significantly impair the patient’s quality of life (QoL) depending on its location. The progressive nature of SFVMs often leads to chronic pain, recurrent infections, bleeding episodes and functional impairments, which in turn lead to psychological distress and social isolation. Reduced physical functioning further decreases the overall QoL for affected patients. Currently, the most used treatment modalities include surgical resection and sclerotherapy (injection of a chemical solution directly into the varicose or spider vein) to close and destroy the abnormal veins. Repeated surgical procedures are often paired with sclerosis therapy, which is less invasive but requires multiple sessions and doesn’t necessarily provide lasting relief. Because of these drawbacks, scientists are actively pursuing novel therapeutic approaches with the goal of incorporating the best of the novel small molecules that have already been developed for conditions like cancer and inflammation (Wirth and Sundine 2007).

The Vascular Anomaly-sirolimus-Europe (VASE) trial, is a phase III trial of sirolimus for the treatment of SFVMs (Seront et al. 2023). Preliminary findings from the VASE trial show that 85 per cent of patients either completely or some-what improve their clinical symptoms within the first month of treatment, and sirolimus may allow surgeons and sclerotherapists to treat unsuitable patients currently deemed ‘too sick’ to consider an intervention. There have been some unavoidable grade 3-4 adverse events, but most were reported to be infections, arterial thromboses/embolisms, stenosis of peripheral vessels, peripheral nerve injury, and hypertension – all events that can be managed or resolved with routine care and discontinuation of the medication (Seront et al. 2023). We shall, in this and the next two sections, categorize the different types of abnormalities, assess the VASE trial dataset its structure and features. We then analyze various endpoints using the classical technique of a linear mixed model and highlight the limitation of this approach when it comes to endpoint selection within this specific context. We finally analyze the trial endpoints again

using the GPC approach, both matched and unmatched.

6.2 Classification of various anomalies

This section gives only a short list of the different anomalies associated with SFVMs, the section having been taken from Wirth and Sundine (2007). Readers can go to their article for a thorough explanation of each anomaly.

6.2.1 Lymphatic Malformations

Lymphatic anomalies are malformations of lymphatic channels, which may be fluid-filled cysts (macrocytic) or more diffuse lesions (microcystic), and occasionally contain both lymph and blood vessels (combined with blood vessels). These lesions often feed into each other and are often part of a larger malformation that can involve multiple regions of the body. Lymphatic malformations have a preference for the head and neck, but they can also affect the arms and legs as well as the trunk. These lesions develop early in life and increase in size along with the child. Microcystic lymphatic malformations commonly appear as clear plaques or diffuse lesions on the skin, or as reddish papules or hairy patches. But they can also arise inside the body in the mucous membranes or in internal organs. Macrocytic malformations appear as soft, multilobular spherical masses, often bluish in colour. Among the complications of lymphatic malformations are intralesional bleeding and infection. Bleeding causes an acute increase in volume with pain. Infection may require antibiotics or even hospitalisation. Rapid growth sometimes accompanies respiratory infections and can result in obstructive airway disease causing respiratory arrest or death in extreme cases. The management of lymphatic malformations ranges from simple observation to surgical excision. Large limb lymphatic malformations may respond to compression therapy. Multiple repeated aspiration and injections of sclerosing agents can be useful in the management of very large cysts. The possibility of surgical removal offers the only therapeutic possibility for a cure, although this is limited by anatomical constraints. But there is a risk for a recurrence and complications can arise.

6.2.2 Venous Malformation

Venous malformations (VMs) are low-flow vascular lesions present at birth; soft, bluish lesions, they can be single or multiple, and various in size. These vein-like structures have thin, large-bore, dilated venous channels that can increase in size depending on pressure or during puberty. Symptoms depend on the size and location of a VMs. The symptoms range from pain, swelling and functional impairments such as facial asymmetry, dental problems, and airway obstruction. Diagnostic tools include clinical examination, X-rays, ultrasound, MRI and CT scan, and treatment can include sclerotherapy or surgery. VMs can be associated with genetic syndromes like Maffucci's syndrome and blue rubber bleb nevus syndrome.

6.2.2.1 Maffucci's syndrome

Maffucci's syndrome, named after the pathologist Angelo Maffucci, who described it in 1881, is an extremely rare and non-inherited syndrome of venous malformations (lesions composed of aberrant veins) and a condition called enchondromatosis (many little bone tumors consisting of cartilage develop in the bones). The condition appears equally in men

and women. The lesions are present at birth and continue to grow into adulthood. The enchondromas are predominantly in the tubular long bones of the arms and legs, leading skeletal deformities and greater risk of fracture. Dwarfism may also occur. Enchondromas rarely affect the skull or pelvis but for the risk of malignant transformation is up to 30 per cent, therefore any painful lesions ought to be imaged with MRI and considered for a biopsy. Small venous malformations of this syndrome form red-blue or blue nodules on the skin, which can be slightly compressed, just like in Bean's syndrome.

6.2.2.2 Blue Rubber Bleb Nevus Syndrome of Bean

William Bennett Bean introduced the term blue rubber bleb nevus syndrome (BRBNS) in 1958 to describe the case of an elderly man affected by venous malformations (abnormal veins) in the skin and different parts of the body. The skin lesions are soft, bluish nodules, sometimes painful and usually associated with some small, colourless, dome-shaped lesions, which Greenberg called 'rubber blebs'. Visceral lesions can be easily seen in the gut during endoscopic examination and are the cause of recurrent bleeding, often resulting in anaemia. BRBNS can also affect many other organs, from the mouth, anus, bladder, brain, spinal cord, liver, spleen, lungs, bones and muscles. Surgical excision, cryosurgery (freezing) or laser therapy are the possibilities for treatment.

6.2.3 Capillary Malformations

Capillary malformations (CMs) are another type of slow-flow vascular lesion. CM may appear as minor, isolated macules, but at other times there may be extensive involvement of large areas of the body, with significant overgrowth. Although a CM is often a purely cosmetic problem, it can be associated with underlying abnormalities such as spinal dysraphism, and, if left untreated, it is prone to grow and result in soft tissue and skeletal overgrowth. Facial CMs can affect one or both sides of the face and one or more areas supplied by the trigeminal facial nerve. Extremity CMs generally grow with the patient. Bony overgrowth can lead to crooked or missing teeth, for which patients may require laser therapy to lighten the lesion and surgical intervention to correct any deformities. A widely used but controversial treatment involves the flashlamp pulsed dye laser, which penetrates only superficially and might not be effective for deeper lesions. Laser therapy in general runs the risk of ulceration, scarring, and pigment change, and other types are available such as Nd:YAG and carbon dioxide lasers.

6.2.4 Conclusion

SFVMs are a heterogeneous group of disorders arising from vascular developmental errors. Clinical manifestations and complications of SFVMs are different for the three main types of malformations: capillary, venous and lymphatic malformations. SFVMs can cause pain and functional disabilities, such as facial asymmetry or leg length discrepancies, swelling, psychological problems and social limitations, and recurrent infections. Because of the various symptoms, we see that the choice of a single primal endpoint might be a challenge.

These SFVMs occur at birth, although it may not always be obvious that they are present. Because of their chronic, progressive nature, the symptoms of these malformations often change with time – particularly during hormonal changes, when they can evolve and become even more pathological – and can become aggravated with trauma or infection.

The heterogenous nature of SFVMs with their progressive nature and complications underline the need of treatment strategies for them. Though surgical excision and sclerotherapy has a very limited role, most recent advancement in sirolimus has opened a new horizon of hope for the treatment of SFVMs.

6.3 Analysis of the data

The dataset from the VASE trial includes information on both categorical and numerical data variables like patient demographics, clinical features, and outcomes of treatment. The trial’s primary outcome measures were patients’ global self-assessment of evolution (overall perception of whether they felt better), pain (continuous and episodic), functional limitation (perceived level of disability), and lesion size. Pain and functional limitation were assessed through a visual analogue scale (VAS) ranging from 0 (no pain/limitation) to 10 (excessive pain/limitation) and frequency of exacerbations of pain were recorded. Exploratory analyses will focus on pain (permanent and occasional) and functional limitation, and their frequency of occurrence and intensity. The clinical data are longitudinal, with assessments taken at 1, 2, 3, 6, 9, 12,15,18,21 and 24 months after enrolment into the study and initiation of the sirolimus treatment. Note that for this analysis we will restrict ourself until the 12-month (follow-up 6). Because we will assume that we don’t have missing data up to the 12-month and will interpret the missing value to indicate that a symptom or event wasn’t present. For example, if pain is missing at a given time point it will be assumed that the patient did not have any pain. We made this assumption based on information provided by those responsible for the data, who confirmed that this assumption was accurate. According to the trial objective, we are going to recode the diagnostic groups into broader categories based on type of vascular malformation:

- Group 1: unifocal and multifocal venous malformations (VM), blue rubber bleb nevus syndrome (BRBNS), capillary-venous malformations (CVM), Capillary malformation with overgrowth (CMO), Capillary malformation with dilated veins (CMDV), Diffuse capillary malformation with overgrowth (DCMO), Verruqueous venous malformation (VVM)
- Group 2: Lymphatic malformations (LM) and lymphaticovenous malformations (LVM).
- Group 3: Combined vascular malformations (CLVM), Klippel-Trenaunay syndrome (KTS), and CLOVES syndrome.
- Group 4: Gorham-Stout disease (GSD) and generalized lymphatic anomaly (GLA).
- Group 5: PTEN hamartoma tumor syndrome (PHTS).
- Group 6: Not diagnosed.

The distribution of participants (132 in total) across the six groups is as follows: 87 in group 1, 22 in group 2, 11 in group 3, 8 in group 4, 3 in group 5, and 1 in group 6.

In the remaining sections, we present exploratory analyses of the main key variables of interest, their distributions, evolution over time, and the associations with treatment patterns and patient outcomes.

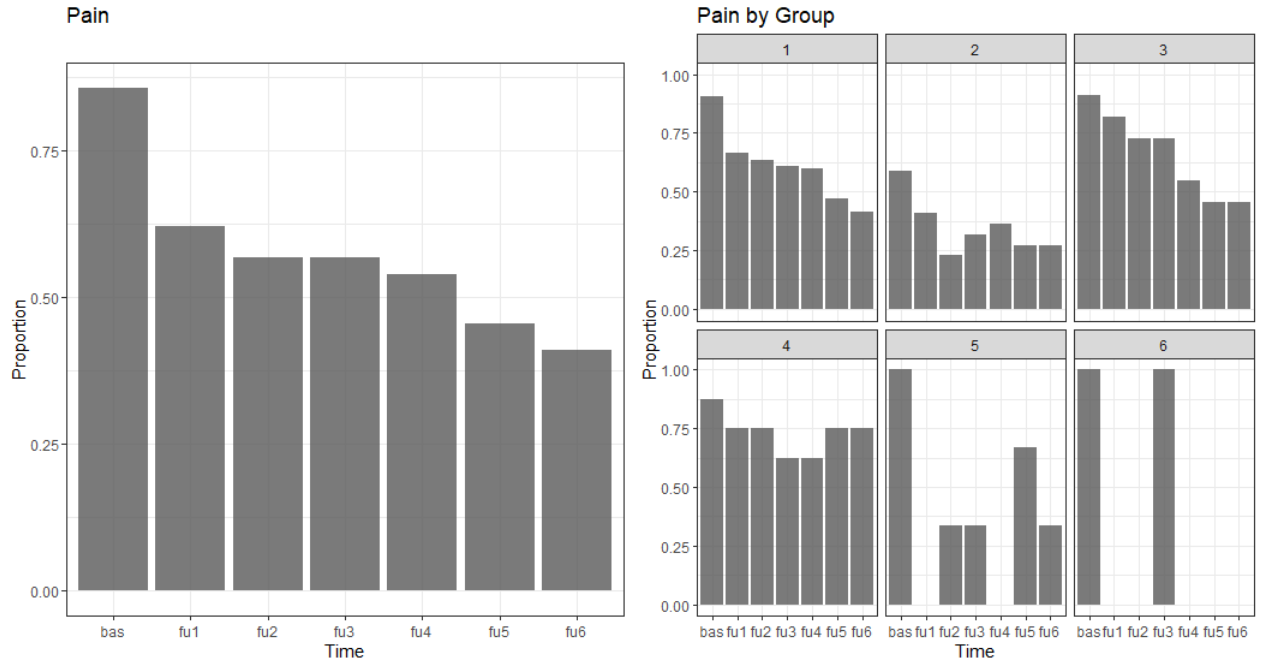


Figure 6.1: The left bar plot displays the proportion of individuals reporting pain at each follow-up. The right bar plot shows the proportion of individuals reporting pain, categorized by their diagnostic groups.

6.3.1 Exploratory analysis

We started our exploratory analysis by plotting the percentage of people and the percentage of people by group who mentioned to suffer from pain (Figure 6.1), pain exacerbation (Figure 6.2) and functional limitation (Figure 6.3).

The figures reveal an overall reduction in the proportion of individuals reporting pain, with an approximate 40% decrease from baseline to the 12-month follow-up (follow-up 6). This decrease is observed across all groups, although it is less pronounced in group 4. In group 6, consisting of undiagnosed individuals, the single participant reported pain at baseline and at the 3-month follow-up.

A similar trend is evident in pain exacerbation, with roughly a 50% reduction in reported occurrences from baseline to follow-up 6. This reduction is consistent across most groups, except for group 4, where no decrease is observed between baseline and follow-up 6. The single participant in group 6 reported pain exacerbation at baseline and the 3-month follow-up.

Regarding functional limitations, a pattern similar to the previous variables emerges. A slight decrease in the percentage of individuals reporting functional limitations is observed in group 4 from baseline to follow-up 6. Again, the participant in group 6 reported experiencing functional limitations at baseline and the 3-month follow-up.

For each variable (pain frequency, pain intensity, pain exacerbation frequency and severity, and functional limitation frequency and severity), we plotted the mean score over time (baseline to follow-up 6) and observed substantial reductions from baseline to the final follow-up (See Appendix B).

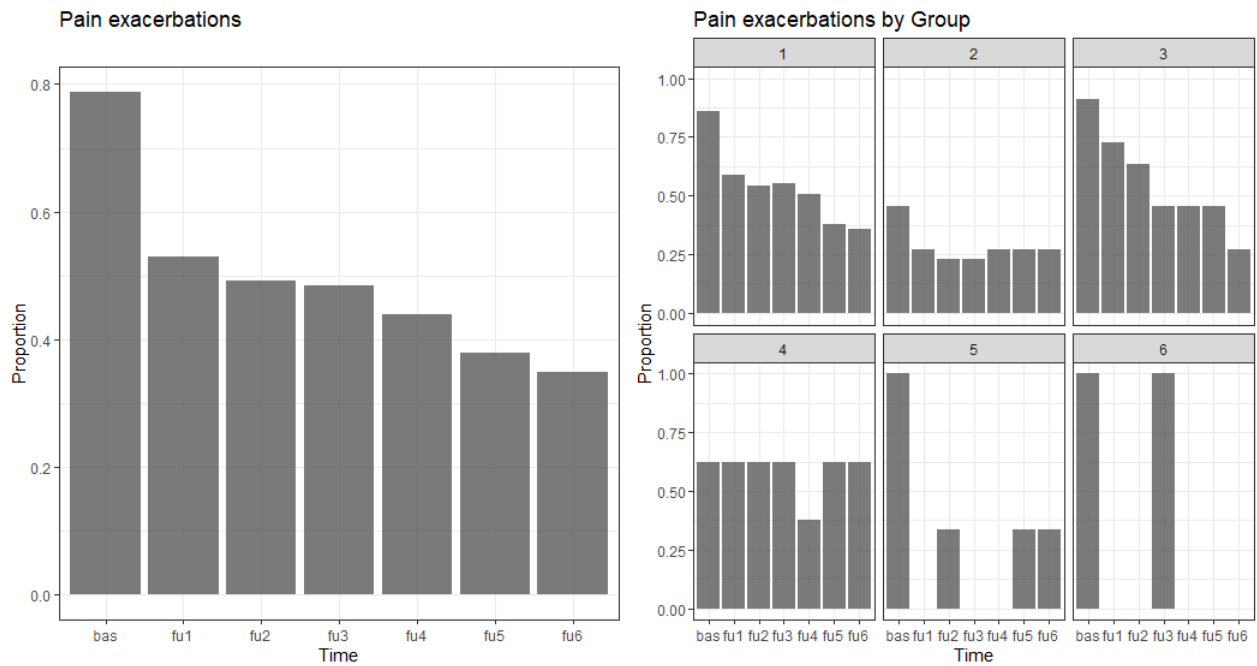


Figure 6.2: The left bar plot displays the proportion of individuals reporting pain exacerbations at each follow-up. The right bar plot shows the proportion of individuals reporting pain exacerbations, categorized by their diagnostic groups.

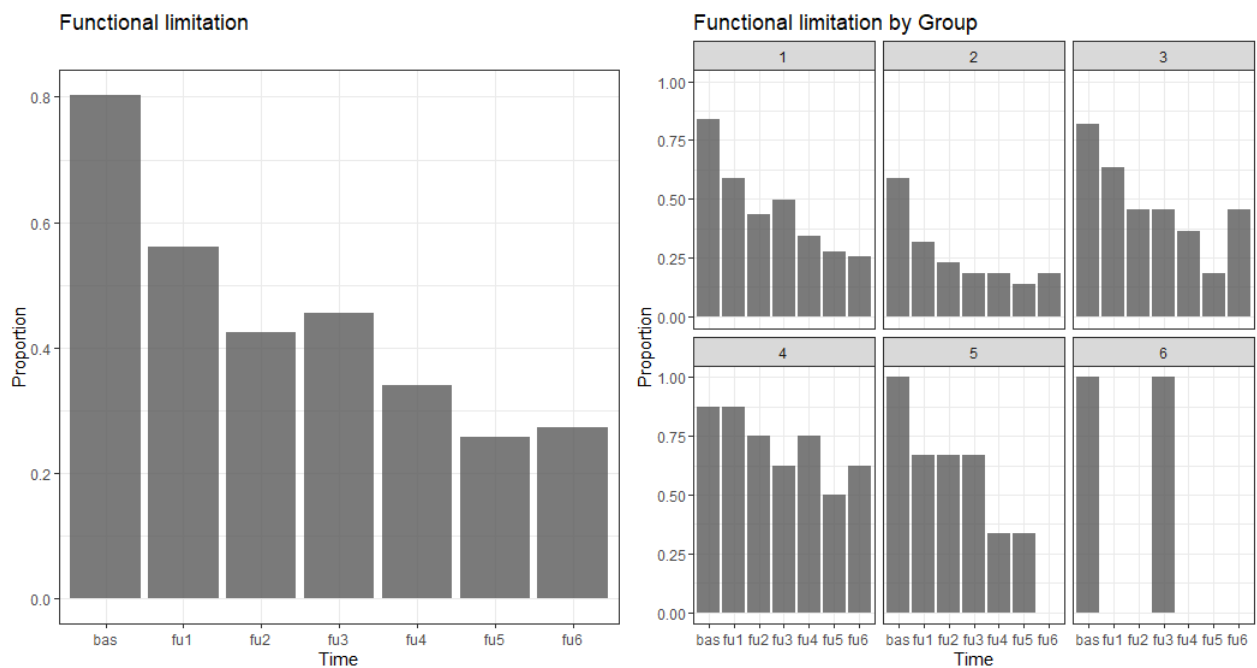


Figure 6.3: The left bar plot displays the proportion of individuals reporting functional limitations at each follow-up. The right bar plot shows the proportion of individuals reporting functional limitations, categorized by their diagnostic groups.

6.3.1.1 Overall Results

The most pronounced decreases typically occurred between baseline and the first follow-up. While the rate of improvement sometimes slowed down, the downward trend generally persisted throughout the study period. Notably, the error bars (representing standard error) generally did not overlap between consecutive time points, suggesting that the observed decreases were statistically significant. Detailed plots can be found in the Appendix.

6.3.1.2 Group-Specific Findings

Across the six groups, most reproduced the overall trends of improvement in pain-related outcomes. Group 1, with the largest sample size, consistently showed substantial decreases in all measures. Group 2 exhibited similar patterns but with slightly smaller effect sizes. Group 3 generally followed the trends, although the improvements in pain exacerbation severity and functional limitations were less pronounced than in other groups. Group 4 displayed more variability, particularly in pain exacerbation frequency, possibly due to its smaller sample size. The small sample sizes of Groups 5 and 6 (3 and 1 participant, respectively) limit the conclusions that can be drawn; however, Group 5 notably showed a temporary increase in functional limitations at follow-up 5.

6.3.2 Linear-mixed effect model

The longitudinal nature of the sirolimus trial data lends itself naturally to analysis using linear mixed-effects models with random intercepts and slopes. Since there is no control arm in the trial, we cannot directly assess the treatment effect of sirolimus using a binary indicator variable. Instead, we focus on the fixed effect of time (in months) as a proxy for treatment effect. For instance, if the outcome of interest is pain frequency (a variable taking value between 0 and 10), a negative and statistically significant fixed effect of time would suggest that sirolimus decreases pain severity over time, thus indicating treatment effectiveness.

However, a key limitation of linear mixed-effects models is the need to choose a single outcome for analysis. Let's illustrate this with the example of pain frequency as our outcome of interest. In this case, the random intercept and random slope model with time as the covariate can be formulated as follows:

$$Y_{ij} = \alpha_0 + \beta_{0i} + (\alpha_1 + \beta_{1i})t_{ij} + \epsilon_{ij}.$$

Where :

- Y_{ij} is the value of severity of pain exacerbations of patient i at time (month) $j \in \{0, 1, 2, 3, 6, 9, 12\}$
- α_0 is the average intercept
- α_1 is the average slope (fixed effect) of time
- β_{0i} is the random deviation in the intercept
- β_{1i} is the random deviation in the slope
- ϵ_{ij} is the random error term

Table 6.1: Estimated fixed effects and variance components for the random intercept and random slope model of frequency of pain, with time (in months) as the covariate

Fixed effect :			
	Estimate	Std.Error	t value
Intercept	0.827	0.052	15.623
month	-0.0299	0.005	-5.834
Parameter of variance :			
Group	Variable	Variance	Corr
ID	Intercept	0.232	
	Month	0.0009	-0.44

Table 6.2: Estimated fixed effects and variance components for the random intercept and random slope model of functional limitation severity, with time (in months) as the covariate

Fixed effect :			
	Estimate	Std.Error	t value
Intercept	3.212	0.216	14.85
month	-0.221	0.018	-11.90
Parameter of variance :			
Group	Variable	Variance	Corr
ID	Intercept	4.809	
	Month	0.011	-0.92

We used R and the lme4 package to fit the linear mixed-effects model. The results are presented in Table 6.1, with the adjusted model illustrated in Figure 6.4.

Table 6.1 indicates that the fixed effect of time (month) is -0.0299 . The corresponding 95% confidence interval for this slope parameter is $[-0.04, -0.019]$. Since this confidence interval does not include zero, we can conclude with 95% confidence that there is a statistically significant decrease in pain frequency over time in patients receiving sirolimus treatment. This finding suggests that sirolimus is effective in reducing the frequency of pain.

While we observed a statistically significant reduction in the frequency of pain exacerbations (Table 6.1, Figure 6.4), the effect size appears small, raising questions about its clinical relevance.

In contrast, when considering the severity of functional limitations as the primary outcome, the estimated fixed effect of time (month) is also negative and statistically significant (Table 6.2). Notably, the effect size is approximately six times larger than that observed for pain exacerbation severity, suggesting maybe a clinically relevant improvement in functional limitations with sirolimus treatment.

In the context of rare diseases, where defining a single primary endpoint can be challenging, the limitations of the linear mixed modeling approach become apparent. Choosing the wrong endpoint as primary may lead to the erroneous conclusion of no treatment effect or negative effect, even when a clinical relevant effect exists on another outcome. Nevertheless, the available data does not indicate any harmful effects of sirolimus on the studied patient population.

Furthermore, another difficulty we encountered with linear mixed-effects models was the issue of convergence. For most outcomes, we were unable to achieve model convergence. This is a serious limitation, as designing a study around a specific endpoint analysis that

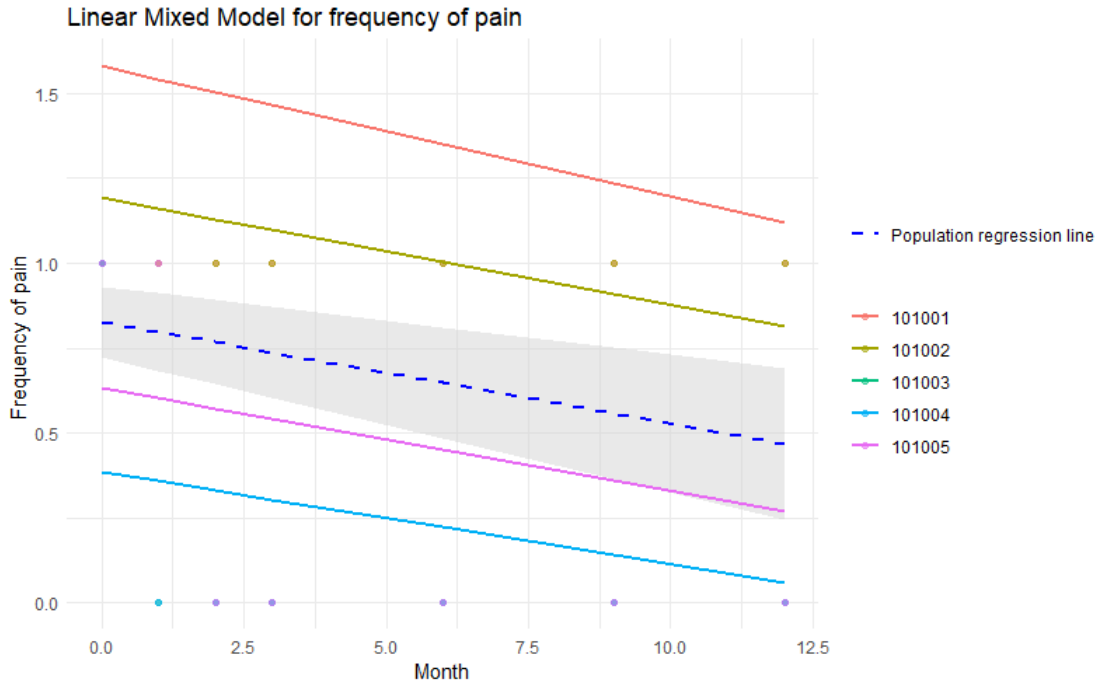


Figure 6.4: Visual representation of the linear mixed model for pain frequency over time (months). Individual participant trajectories (colored lines) are shown alongside the population average trend (blue dashed line) and its 95% confidence interval (grey shaded area).

ultimately cannot be utilized raises questions about the appropriateness of this method for such data. If convergence cannot be achieved, the validity of the results is compromised, blocking our ability to draw reliable conclusions from the trial data. The reader can use the Shiny application (see Section 6.3.4) that we have developed in order to explore those issue in more details for various clinical endpoint.

6.3.3 Generalized pairwise comparison

In Section 6.3.2, we highlighted the challenges of selecting a single primary endpoint for analysis. The GPC methodology addresses this issue by enabling comparisons between two groups of observations based on multiple, hierarchically ordered endpoints. Although the GPC framework effectively incorporates various outcomes, it also presents certain limitations. A crucial initial consideration is whether to use a matched or unmatched approach for the analysis. During our simulation (see Section 5.3.2), we observed that when no inter individual correlation between measurements is expected, the unmatched method generally outperformed the matched method in term of power. However, as the inter individual correlation between measurements increased, the difference in performance between the two methods became less pronounced. While no clear superiority in terms of power was observed for the matched method, it offered the advantage of being more interpretable.

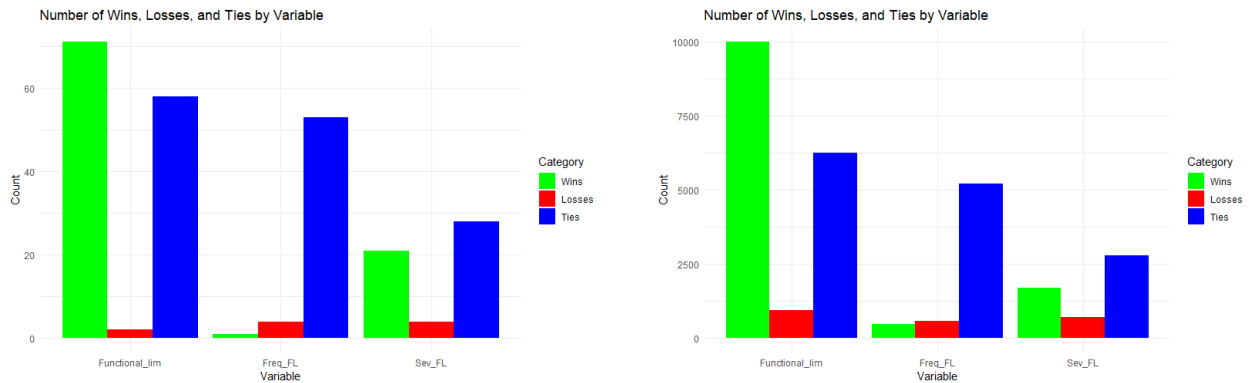
Indeed, the two methods measure different parameters. As a reminder, the matched method measures Γ , the probability that an individual fares better under the treatment (sirolimus in our application) than under control (baseline). The unmatched method, on the other hand, measures Δ , the expected improvement in terms of quantiles across the entire population (baseline and sirolimus) for a patient receiving sirolimus. These two metrics can lead to different conclusions (Fay et al. 2018), but in the case of sirolimus data we did not

find a setting where it was the case. Both methods lead to same conclusion.

6.3.3.1 Matched vs. unmatched GPC Analysis of the sirolimus Trial

By using the Shiny application that we have developed for the sirolimus data, we can explore how the approaches might differ with each other. For example, if we decide that the three most important outcomes of the trial are the variable functional limitation a binary variable (does the patient experience functional limitation), frequency of functional limitation (how often does the patient functional limitation) and finally the severity of functional limitation. If the threshold for the three outcomes is set to 0, meaning that any reduction is considered as relevant. The matched and the unmatched methods do agree. For the matched methods, we obtain an estimation of the net benefit ($\hat{\gamma}$) which is equal to 0.633 and p-value which is less than 0.001. So, from the data we have strong evidence that the true net benefit (Γ) is different from 0 and we would conclude that the treatment under sirolimus is beneficial. Similarly, the unmatched methods give us a net benefit estimate ($\hat{\delta}$) of 0.578 and a p-value less than 0.001. We would therefore reject the null hypothesis that the true net benefit (Δ) is different from 0 as well.

Figure 6.5 displays the distribution of pairs for the matched GPC method (left) and the unmatched method (right) applied to the sirolimus trial data.



(a) Distribution of pairs for the matched GPC analysis of the sirolimus trial, prioritizing functional limitation, functional limitation frequency, and functional limitation severity as outcomes.

(b) Distribution of pairs for the Unmatched GPC analysis of the sirolimus trial, prioritizing functional limitation, functional limitation frequency, and functional limitation severity as outcomes.

Figure 6.5: Comparison of the distribution of pairs for the matched vs. unmatched GPC analysis of the sirolimus trial prioritizing functional limitation, functional limitation frequency, and functional limitation severity as outcomes.

Even though the GPC method offers the major advantage of incorporating multiple endpoints, it lacks the ability to be applied in a longitudinal context. The linear mixed model presented in Section 6.3.2 utilizes all available information for each patient, whereas the GPC methodology, in our case, only considers the values at baseline and at the end of the 12-month treatment period.

In our application, this limitation of GPC may not significantly impact our findings. For example, applying the matched GPC method to the pain exacerbation severity outcome yields an estimated net benefit of 0.6439 (p-value < 0.01), leading to the conclusion that sirolimus treatment effectively reduces pain exacerbation severity, similar to the results obtained with the linear mixed model. However, this conclusion relies on the strong assumption of

complete data up to the 12-month follow-up, with missing values assumed to represent no pain exacerbation.

Verbeeck (2022) proposed a GPC methodology for longitudinal data by prioritizing time points and outcomes based on clinical importance. In our case, the presence of pain (Yes/No) at follow-up 6 would be prioritized over pain severity at follow-up 6. Ties would be resolved by considering the presence of pain at follow-up 5, then pain severity at follow-up 5, and so on. While this approach is straightforward, it raises some concerns.

For instance, if many patients discontinue the trial due to treatment toxicity despite initial effectiveness, this methodology, by prioritizing earlier time points where the treatment might have appeared beneficial, could still indicate a positive overall treatment effect. Similarly, if a treatment is effective in the short or medium term but loses efficacy in the long run, the prioritization scheme may result in many ties when comparing baseline to the last follow-up. However, since earlier time points would be taken into account, the methodology might identify some instances where the treatment was effective, leading to the erroneous conclusion of sustained treatment efficacy, even though the long-term effect is actually neutral or negative.

6.3.4 Application Shiny

The Shiny application we developed can be accessed through the following link: <https://kamalgasser.shinyapps.io/SirolimusApp/>. The application has three main tabs: 'Load the Data,' 'Linear Mixed Model,' and 'GPC.' The first tab, 'Load the Data,' allows the user to upload the dataset using the 'Browse' button. The dataset must be in the exact format provided for this thesis by my supervisor, Prof. Catherine Legrand. The user can then select the variables of interest. So far, we have included the following options: 'Pain,' 'Frequency of Pain,' 'Severity of Pain,' 'Pain Exacerbation,' 'Frequency of Pain Exacerbation,' 'Severity of Pain Exacerbation,' 'Functional Limitation,' 'Frequency of Functional Limitation,' and 'Severity of Functional Limitation.' The user can also specify the follow-up period they wish to investigate. Although the application includes data up to follow-up 9 to allow for broader use, we recommend limiting the analysis to follow-up 6, as the app is currently optimized to function correctly only up to that point. Figure 6.6 shows the main button of the first tab. Once the user has loaded the data, selected the variables of interest and the follow-up, he will see on the right panel the head of the dataset and below a summary table of the variables by follow-up.

In the 'Linear Mixed Model' tab, the user can select the variable for which they want to fit a linear mixed model with random intercept and random slope. Note that only continuous variables can be used: 'Frequency of Pain,' 'Severity of Pain,' 'Frequency of Pain Exacerbation,' 'Severity of Pain Exacerbation,' 'Frequency of Functional Limitation,' and 'Severity of Functional Limitation.' To choose any of these variables, they must first be selected in the 'Load the Data' tab. The user can then click on the 'Fit Model' button, and the right panel will display the output from the lme4 package, along with a graph of the regression results below it (see Figure 6.7). Note that the regression will go maximum until follow-up 6 (12 months).

Finally, the 'GPC' tab (see Figure 6.8) allows the user to select the variables for a GPC analysis. The order of the variables is crucial, as GPC prioritizes them with the first variable being the most important, followed by the others in sequence. The user can also specify the significance level for inference about the net benefit and set a threshold of clinical significance for each selected variable. Note that the GPC tab functions correctly only up to follow-up 6.

Shiny app for the sirolimus trial

Load the data [Linear mixed model](#) [GPC](#)

Choose R Data File

Browse... No file selected

Data Format

☒ RData
☐ RDS

Variables

Pain

Follow-up

1

Quit

Figure 6.6: Tab 'Load the data' of the sirolimus Shiny application

Shiny app for the sirolimus trial

[Load the data](#) [Linear mixed model](#) [GPC](#)

Select response variable

Sev_pain

Fit Model **Quit**

```
Linear mixed model fit by REML ['lmerMod']
Formula: Sev_pain ~ month + (1 + month | record_id)
Data: data
Control: lmerControl(optimizer = "optimx", optCtrl = list(method = "nlopt"))

REML criterion at convergence: 4030.1

Scaled residuals:
    Min       1Q   Median       3Q      Max
-2.8554 -0.5187 -0.1628  0.3615  3.7299

Random effects:
 Groups   Name      Variance Std.Dev. Corr
record_id (Intercept) 4.40336  2.0984
          month       0.01406  0.1186  -0.77
Residual    3.30149  1.8170
Number of obs: 924, groups: record_id, 132

Fixed effects:
              Estimate Std. Error t value
(Intercept)  2.77033    0.20393   13.59
month        -0.18062    0.01777  -10.16

Correlation of Fixed Effects:
      (Intr)
month -0.675

Linear Mixed Model for Sev_pain
```

Figure 6.7: Tab 'Linear mixed model' of the sirolimus Shiny application

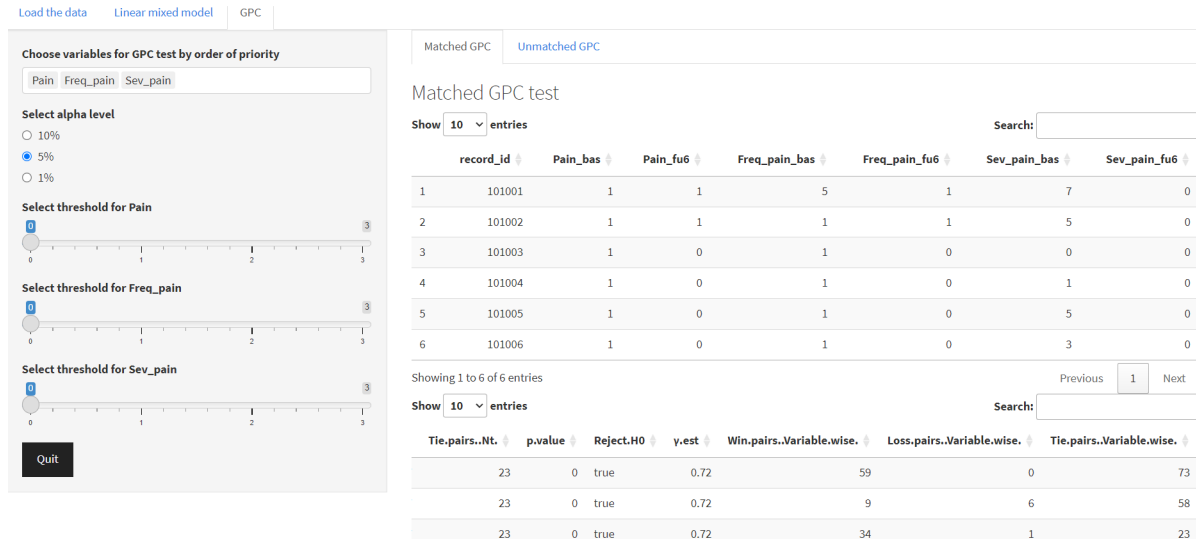


Figure 6.8: Tab 'GPC' of the sirolimus Shiny application

The right panel contains two sub-tabs: one for the matched method and one for the unmatched method. Each sub-tab displays the head of the dataframe in wide format. Below this, a table presents the total number of wins, losses, and ties, both overall and for each outcome. The estimated net benefit and the corresponding p-value are also displayed. Below this table, a graph shows the distribution of pairs by outcome. An important remark is that the inference technique used for the unmatched method is the asymptotic Munzel test.

The development of a fully optimized application could be a substantial project on its own. We developed this tool rapidly, with the primary goal of investigating, up to follow-up 6, the different methods presented in this thesis for analyzing the sirolimus trial. While the app serves its intended purpose for this analysis, certain design enhancements and additional functionalities are still needed for broader application.

7. Conclusion

In this thesis, we have examined the multiple challenges faced by rare disease trials, as identified by Kandi and Vadakedath (2023):

- Limited number of available patients.
- Selection of the primary endpoint.
- Specific trial designs that may lack a control arm, such as those with only baseline and end-of-treatment period values.

The GPC methodology offers a significant advantage in rare disease trials by allowing the inclusion of multiple endpoints to evaluate treatment effects comprehensively. This approach enables a global assessment of treatment impact on disease symptoms without the need for individual outcome testing and multiplicity correction (Deltuvaite-Thomas et al. 2023).

In designs where each patient receives the new intervention (treatment), and the control group consists of their baseline values, we can utilize either the unmatched or matched GPC methods. These two methodologies estimate different parameters. The unmatched method assumes two groups: the control group (baseline values) and the treatment group (end-of-treatment values). This method estimates the net benefit (Δ) by comparing every pair between the two groups, measuring the expected improvement for a patient (Fay et al. 2018). As the two groups are not independent, we demonstrate in Section 5 that the (studentized) Munzel test should be used for proper inference about the parameter.

Conversely, the matched method compares pairs at the individual level, estimating Γ , the probability that a patient performs better under treatment than the standard care. This metric is easier to interpret and communicate, and for inference, we use the methodology proposed by Matsouaka (2022). Although these parameters usually align, discrepancies can occur. For example, if a minority of patients show significant improvement while the majority experience moderate degradation, Δ may indicate a positive treatment effect, whereas Γ may suggest a negative effect. This scenario is discussed thoroughly by Fay et al. (2018), known as the "Hand paradox."

We then used linear mixed models to generate data mimicking rare disease trials, enabling us to use simulations to compare the power of inference methods for unmatched versus matched GPC. Our simulations indicate that the unmatched method has higher power when baseline and end-of-treatment values are nearly independent. Otherwise, neither method shows a clear advantage in terms of power. Therefore, we recommend using the matched methodology in designs comparing baseline versus end-of-treatment periods, as it is more likely that in practice there exist correlation between these values and that Γ is easy to communicate. Our simulations also show that increased correlation between the measures reduces power, suggesting that larger sample sizes are required to achieve reasonable power when expecting high correlation between pre/post treatment measurement.

We applied these methodologies to data from the Sirolimus trial, using a Shiny app to explore different scenarios. From a patient-centric point of view, the fact that each patient can choose their order of priority is advantageous. However, designing a clinical trial where the decision maker needs to choose this order before collecting data is challenging (Deltuvaite-Thomas et al. 2023). Leaving this issue aside, the app consistently showed that Sirolimus effectively reduced symptoms in patients with slow-flow vascular anomalies, regardless of the method used.

In Section 6.3.3, we highlight a major disadvantage of the GPC method: it does not account for the longitudinal nature of the data. Although Verbeeck (2022) proposed a method that prioritizes time points, we argue that this may be inappropriate. To our knowledge, GPC has not been applied to longitudinal data. However, Chen et al. (2014) proposed an extension of the Mann–Whitney–Wilcoxon rank sum test to longitudinal regression analysis using functional response models. While this paper is beyond the scope of this thesis it could be a direction for future research to extend GPC to longitudinal data.

A. Simulation plan for the two continuous outcome

Table A.1: Simulation plan for for the two continuous outcome.

S	$\sigma_{\beta_0}^2$	$\sigma_{\beta_1}^2$	$\sigma_{\beta_2}^2$	$\sigma_{\beta_3}^2$	$\sigma_{\beta_0, \beta_1}$	$\sigma_{\beta_0, \beta_2}$	$\sigma_{\beta_0, \beta_3}$	$\sigma_{\beta_1, \beta_2}$	$\sigma_{\beta_1, \beta_3}$	$\sigma_{\beta_2, \beta_3}$	$\sigma_{\epsilon_1}^2$	$\sigma_{\epsilon_2}^2$	$\rho(Y_{i01}, Y_{i11})$	$\rho(Y_{i02}, Y_{i12})$	$\rho(Y_{i01}, Y_{i02})$	$\rho(Y_{i11}, Y_{i12})$
1	0.0001	1	0.0001	1	0	0	-0.0001	-0.0001	-0.9998	0	1	1	0	0	0	-0.5
2	0.0001	1	0.0001	1	0	0	0.0001	0.0001	0.9998	0	1	1	0	0	0	0.5
3	0.0001	1	5	5	0	0	-0.0004	-2.230	-2.230	4.95	1	0.01	0	0.99	0	-0.706
4	0.0001	1	5	5	0	0	0.0004	2.230	2.230	4.95	1	0.01	0	0.99	0	0.706
5	5	5	0.0001	1	4.95	0	-2.230	-0.0004	-2.230	0	0.01	1	0.99	0	0	-0.706
6	5	5	0.0001	1	4.95	0	2.230	0.0004	2.230	0	0.01	1	0.99	0	0	0.706
7	5	5	5	5	4.95	0	-0.705	-0.705	-1.396	4.95	0.01	0.01	0.99	0.99	0	-0.140
8	5	5	5	5	4.95	0	0.705	0.705	1.396	4.95	0.01	0.01	0.99	0.99	0	0.140
9	5	5	5	5	4.95	4.51	4.160	4.160	3.727	4.95	0.01	0.01	0.99	0.99	0.9	0.605
10	5	5	5	5	4.95	4.51	4.769	4.769	4.933	4.95	0.01	0.01	0.99	0.99	0.9	0.726

B. Bar plot of various variables from the Sirolimus trial

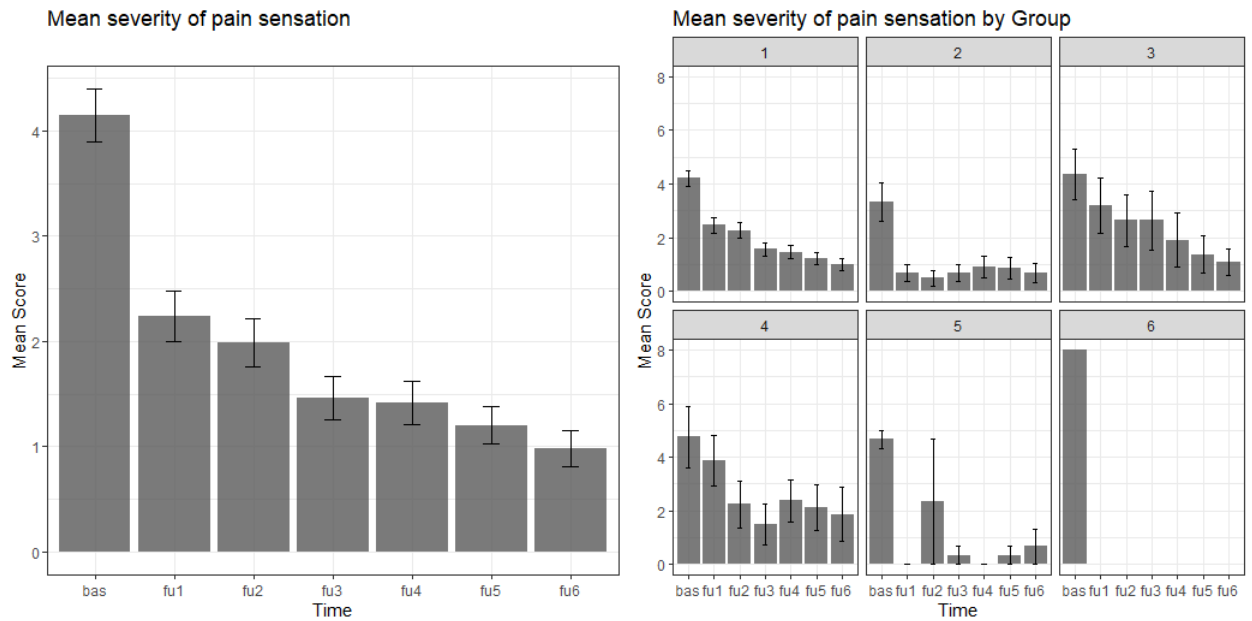


Figure B.1: The left bar plot displays the mean pain severity reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean pain severity reported, categorized by their diagnostic groups.

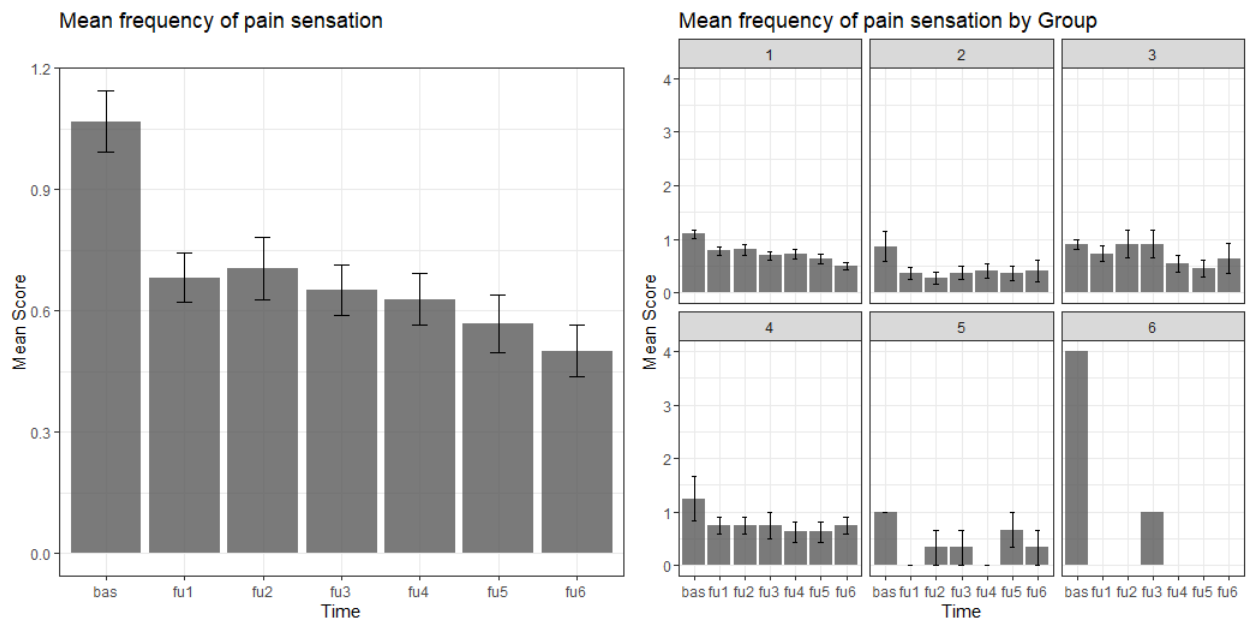


Figure B.2: The left bar plot displays the mean pain frequency reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean pain frequency reported, categorized by their diagnostic groups.

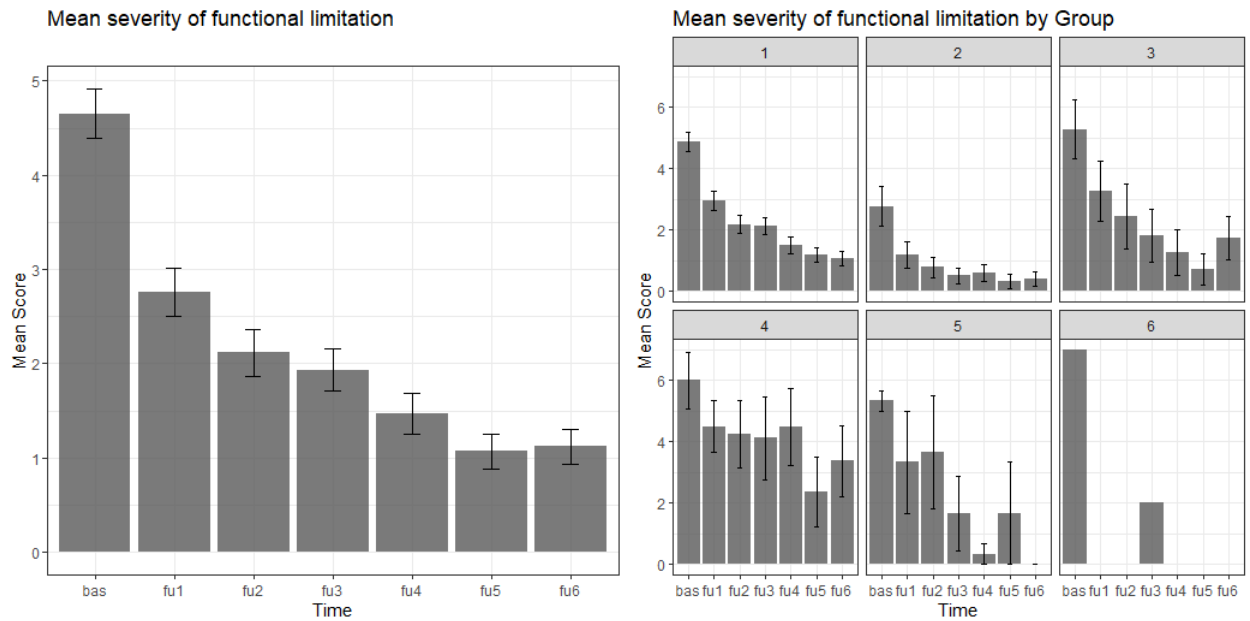


Figure B.3: The left bar plot displays the mean functional limitation frequency reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean functional limitation frequency reported, categorized by their diagnostic groups.

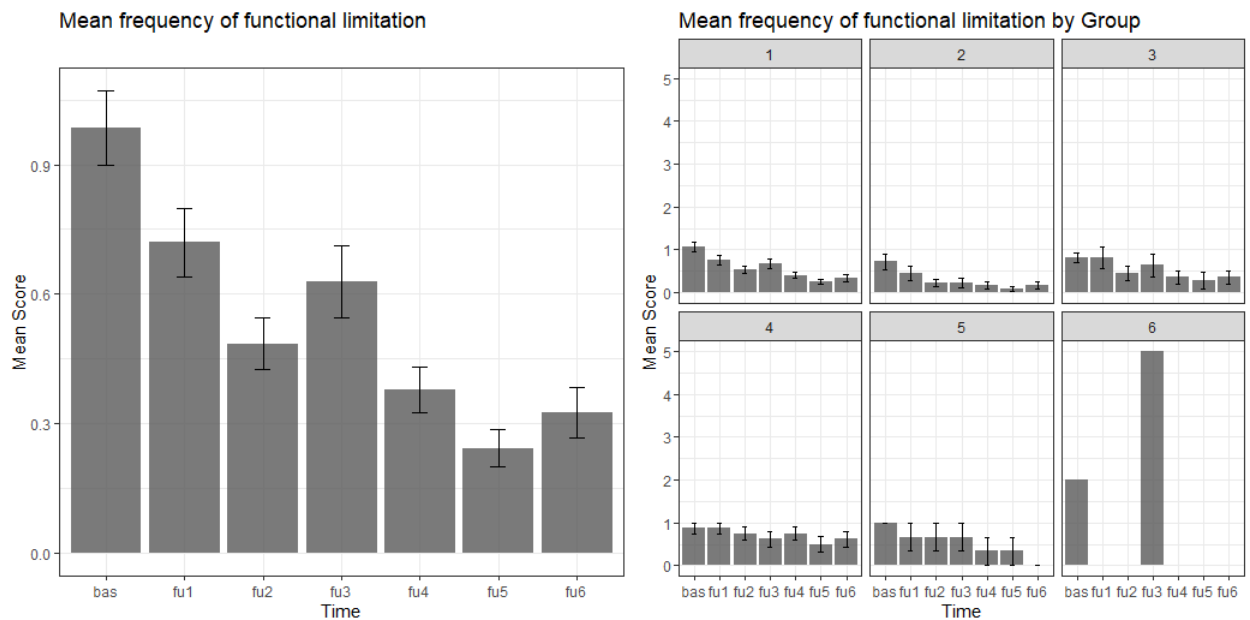


Figure B.4: The left bar plot displays the mean functional limitation frequency reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean functional limitation frequency reported, categorized by their diagnostic groups.

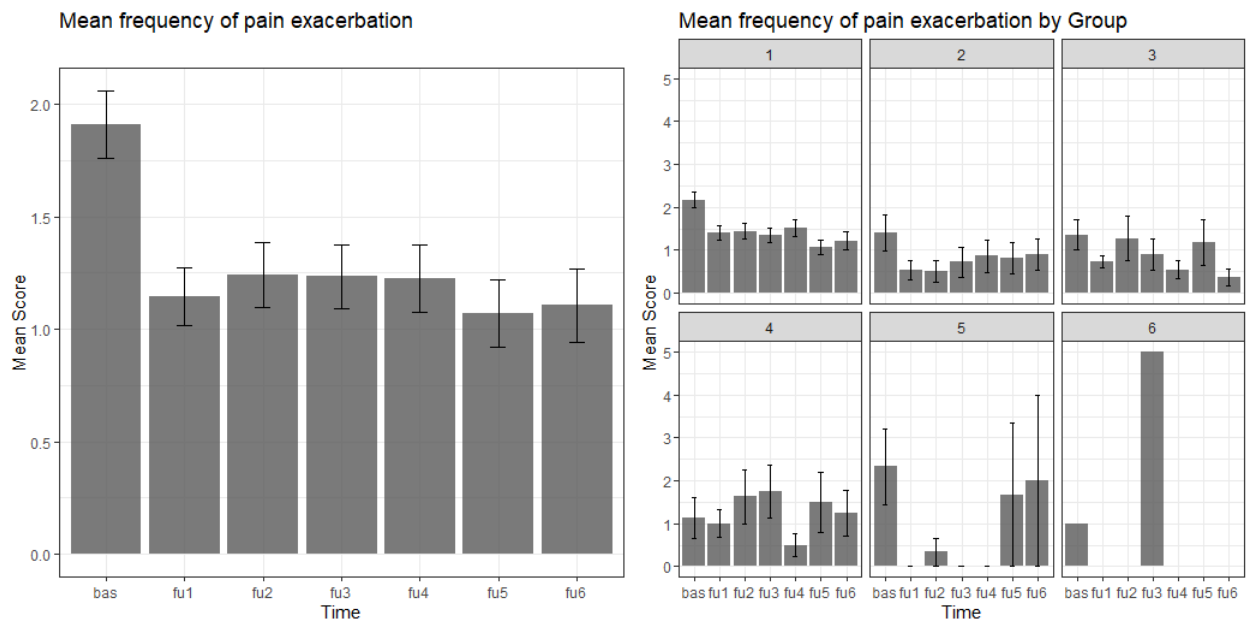


Figure B.5: The left bar plot displays the mean pain exacerbations frequency reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean pain exacerbations frequency reported, categorized by their diagnostic groups.

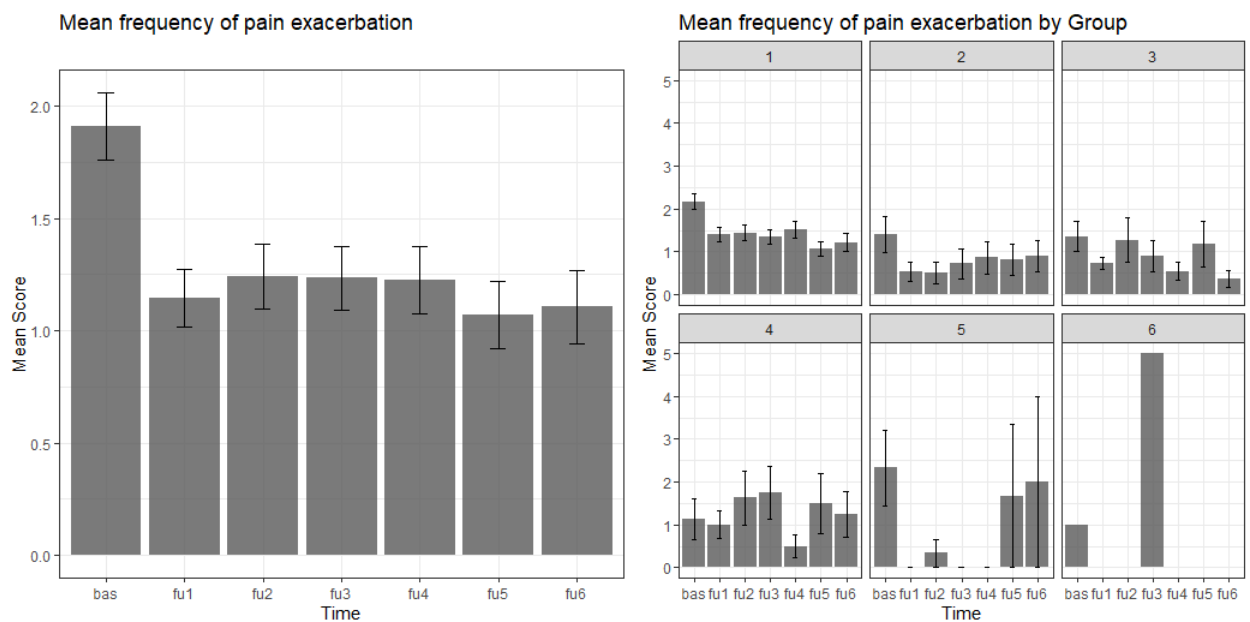


Figure B.6: The left bar plot displays the mean pain exacerbations severity reported at each follow-up with $\pm$ -standard deviation. The right bar plot shows mean pain exacerbations severity reported, categorized by their diagnostic groups.

Bibliography

- Abrahamyan, Lusine, Brian M. Feldman, George Tomlinson, Marie E. Faughnan, Sindhu R. Johnson, Ivan R. Diamond, and Samir Gupta. 2016. “Alternative designs for clinical trials in rare diseases.” *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 172, no. 4 (December): 313–331. ISSN: 1552-4868, 1552-4876, accessed July 12, 2024. <https://doi.org/10.1002/ajmg.c.31533>. <https://onlinelibrary.wiley.com/doi/10.1002/ajmg.c.31533>.
- Brunner, Edgar. 2024. “Relationship Between GPC and Rank Procedures.” Chap. 8 in *Generalized Pairwise Comparisons. A statistical approach to patient-centric medicine*, 175–189.
- Brunner, Edgar, Marc Vandemeulebroecke, and Tobias Mütze. 2021. “Win odds: An adaptation of the win ratio to include ties.” *Statistics in Medicine* 40 (14): 3367–3384.
- Buyse, Marc. 2010. “Generalized pairwise comparisons of prioritized outcomes in the two-sample problem.” *Statistics in medicine* 29 (30): 3245–3257.
- Chen, R, T Chen, N Lu, Hui Zhang, P Wu, C Feng, and XM Tu. 2014. “Extending the Mann–Whitney–Wilcoxon rank sum test to longitudinal regression analysis.” *Journal of Applied Statistics* 41 (12): 2658–2675.
- Deltuvaite-Thomas, Vaiva, and Tomasz Burzykowski. 2022. “Operational characteristics of generalized pairwise comparisons for hierarchically ordered endpoints.” *Pharmaceutical Statistics* 21 (1): 122–132.
- Deltuvaite-Thomas, Vaiva, Mickaël De Backer, Samantha Parker, Marie Deneux, Lynda E Polgreen, Cara O’Neill, Samuel Salvaggio, and Marc Buyse. 2023. “Generalized pairwise comparisons of prioritized outcomes are a powerful and patient-centric analysis of multi-domain scores.” *Orphanet Journal of Rare Diseases* 18 (1): 321.
- Fay, Michael P, Erica H Brittain, Joanna H Shih, Dean A Follmann, and Erin E Gabriel. 2018. “Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments.” *Statistics in Medicine* 37 (20): 2923–2937.
- Friedman, Lawrence M., ed. 2015. *Fundamentals of clinical trials* [in eng]. 5. ed. Cham Heidelberg: Springer. ISBN: 978-3-319-18538-5.
- Giai, Joris. 2021. “Méthodes d’analyse des essais cliniques en médecine personnalisée.” PhD diss., Lyon.

- Hilgers, Ralf-Dieter. 2016. “Design and analysis of clinical trials for small rare disease populations.” *Journal of Rare Diseases Research & Treatment* 1, no. 3 (November): 53–60. ISSN: 25729411, accessed July 12, 2024. <https://doi.org/10.29245/2572-9411/2016/3.1054>. <http://www.rarediseasesjournal.com/articles/design-and-analysis-of-clinical-trials-for-small-rare-disease-populations.html>.
- Kandi, Venkataramana, and Sabitha Vadakedath. 2023. “Clinical Trials and Clinical Research: A Comprehensive Review” [in en]. *Cureus* (February). ISSN: 2168-8184, accessed July 11, 2024. <https://doi.org/10.7759/cureus.35077>. <https://www.cureus.com/articles/128436-clinical-trials-and-clinical-research-a-comprehensive-review>.
- Konietschke, Frank, and Markus Pauly. 2012. “A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data.”
- Legrand, Catherine. 2021. *Advanced survival models*. Chapman / Hall/CRC.
- Matsouaka, Roland A. 2022. “Robust statistical inference for matched win statistics.” *Statistical Methods in Medical Research* 31 (8): 1423–1438.
- Mellerio, Jemima E. 2022. “The challenges of clinical trials in rare diseases” [in en]. *British Journal of Dermatology* 187, no. 4 (October): 453–454. ISSN: 1365-2133, 0007-0963, accessed July 11, 2024. <https://doi.org/10.1111/bjd.21686>. <https://academic.oup.com/bjd/article/187/4/453/6966311>.
- Munzel, Ulrich. 1999. “Nonparametric methods for paired samples.” *Statistica Neerlandica* 53 (3): 277–286.
- Rath, Ana, Valérie Salamon, Sandra Peixoto, Virginie Hivert, Martine Laville, Berenice Segrestin, Edmund AM Neugebauer, Michaela Eikermann, Vittorio Bertele, Silvio Garattini, et al. 2017. “A systematic literature review of evidence-based clinical practice for rare diseases: what are the perceived and real barriers for improving the evidence and how can they be overcome?” *Trials* 18:1–11.
- Senn, Stephen S. 2002. *Cross-over trials in clinical research*. Vol. 5. John Wiley & Sons.
- Seront, Emmanuel, An Van Damme, Catherine Legrand, Annouk Bisdorff-Bresson, Philippe Orcel, Thomas Funck-Brentano, Marie-Antoinette Sevestre, Anne Domp martin, Isabelle Quere, Pascal Brouillard, et al. 2023. “Preliminary results of the European multicentric phase III trial regarding sirolimus in slow-flow vascular malformations.” *JCI insight* 8 (21).
- Verbeeck, Johan. 2022. “(Non)-Prioritizing Generalized Pairwise Comparisons and COVID-19.”
- Wirth, Garrett A, and Michael J Sundine. 2007. “Slow-flow vascular malformations.” *Clinical pediatrics* 46 (2): 109–120.

