Covariance Analysis of Irritable Bowel Syndrome Study II

by **Shintaro Hagiwara**, M.Sc,, and **Patrick Boily**, Ph.D.
Centre for Quantitative Analysis and Decision Support
Carleton University
cqads@carleton.ca

presented to the Canadian College of Naturopathic Medicine

October 2014

This report presents key findings of the covariance analysis that was performed to test the effect of the probiotic agent on the severe sufferers of Irritable Bowel Syndrome (IBS).

It relies in many ways on work previously done by CQADS; as such, large chunks of this report follow the structure and content of "**Covariance Analysis for the 2010 CCNM Pilot Study on Irritable Bowel Syndrome**" [9], a report produced by CQADS in August of 2013.

Background and Executive Summary

Irritable Bowel Syndrome (IBS) is a functional colonic disease with high prevalence. Typical symptoms include "chronic abdominal pain, discomfort, bloating, and alteration of bowel habits" [1]; it has been linked to chronic pain, fatigue, and work absenteeism and is considered to have a severe impact on quality of life [2, 3]. Although there is no known cure for IBS, there are treatments that attempt to relieve symptoms, including dietary adjustments, medication and psychological interventions.

In 2010, the Canadian College of Naturopathic Medicine (CCNM) was commissioned to conduct a pilot study to investigate the effect of a probiotic agent on IBS. The study's details and a preliminary data analysis using hierarchical linear models (HLM) can be found in a preliminary report: it's key findings are that a strong placebo/expectation effect is present in the early stages of the study which is not entirely surprising given the nature of the phenomenon under study, and that there is no strong statistical evidence to suspect that the agent itself has much of an effect on mild to moderate IBS [4]. Furthermore, the key findings from covariance analyses (ANCOVA) on the above data conducted by the Centre for Quantitative Analysis and Decision Support (CQADS) aligned with the analysis using HLM [4,9]; the main ANCOVA results are summarized in the table below.

ANCOVAs for IBS and QoL measures (original dataset)			Sample Size	Initial		At 3 months		p-value
				mean	SD	mean	SD	
All subjects	IBS severity	Placebo	57	273.8	73.7	204.0	97.2	0.095 (0.137†)
		Probiotics	59	268.9	76.4	175.3	78.6	
	QoL	Placebo	58	42.0	20.4	33.4	21.0	**0.056**
		Probiotics	59	40.2	18.6	26.4	17.5	
Severe subjects*	IBS severity*	Placebo	16	363.0	57.9	281.4	121.4	(0.049†)
		Probiotics	19	351.0	44.0	206.3	104.5	
	QoL*	Placebo	17	55.8	21.6	50.6	21.8	0.007
		Probiotics	19	48.3	16.1	29.9	18.0	

Due to the small sample size (and because of issues associated with positively determining membership in the severe sufferer category), the analyses marked with a "*" were not endorsed by CQADS, and are provided for completeness. The significance of the treatment is measured by the *p*-value (*p*-values obtained after analysis on the reduced dataset, for which outliers have been removed, are indicated by a "†").

While some of the results looked promising, no statistical evidence for treatment effect was found at the 95% significance level; furthermore, even had evidence been found at that level, design and recruitment issues would have called their practical significance into question [9].

In 2013, CCNM conducted a second study to investigate the effect of a probiotic agent, this time focusing on severe IBS sufferers. Potential participants were considered to be severe IBS sufferers if they had total IBS severity scores of 300 or higher, with the highest possible score being 500. The study sponsor has expressed interest in analyzing this new data using Analysis of Covariance (ANCOVA) in order to determine whether there is a statistically significant difference between the placebo and the probiotic agent.

ANCOVA is a general linear model which evaluates whether the population means of a dependent/response variable (in this case, total IBS severity score, five IBS sub-scores, and a measure of Quality of Life) are equal across levels of a categorical independent variable (in this case, two treatment effects over time), while statistically controlling for the effects of covariates (in this case, the baseline scores). By comparison with the more traditional analysis of variance (ANOVA), ANCOVA can be used to increase the likelihood of finding a significant difference between treatment groups (when one exists) by reducing the within-group error variance.

The main results of the 7 ANCOVAs (for the new data, imputed with Last Observation Carried Forward, see next page) and the 5 IBS sub-scores ANCOVAs (for the original data, imputed with LOCF, below). Detailed explanations are found in the body of the report.

ANCOVA for the 5 IBS sub-scores (original dataset)		Group	Sample Size	Initial		At 3 months		p-value
				mean	SD	mean	SD	
All subjects	Abdominal pain	Placebo	57	45.26	23.50	30.68	24.51	0.106
		Probiotics	59	43.95	22.79	23.49	21.41	
	Abdominal distension	Placebo	57	51.28	22.93	34.18	26.48	0.445
		Probiotics	59	48.35	25.28	30.19	22.25	
	Satisfaction	Placebo	57	67.79	20.89	56.95	23.40	**0.085**
		Probiotics	59	69.60	23.53	50.42	21.38	
	Interference	Placebo	57	65.81	18.63	47.63	21.16	0.158
		Probiotics	59	59.67	18.13	40.14	20.07	
	Frequency	Placebo	57	43.68	24.32	34.56	27.37	0.347
		Probiotics	59	47.37	28.26	31.04	28.78	

As shown in these tables, the ANCOVA of the two clinical trials to study the effect of the probiotic agent on IBS do not reveal a statistically significant treatment effect. That being said, even though we conclude that there is no evidence to differentiate the treatment effect from the placebo effect, there were some instances when the difference in improvements between the two treatment groups (Probiotics over Placebo in the first study, I over K in the second) were nearly significant (e.g., patients' satisfaction with their bowel movement habits in the first study, and their quality of life in both studies, with *p*-values reaching 0.085, 0.056 and 0.061, respectively).

While the *p*-values themselves may look encouraging, the large placebo effect and high fluctuating nature of IBS on a day-to-day basis make it very difficult to control for the uncertainty in the data. Furthermore, it is far from obvious that these results can be generalized to a larger population due to the non-probabilistic nature of samples collected for the clinical trials, as well as the possibility of a self-reporting bias.

ANCOVA for the 7 core analyses (new dataset)		Group	Sample Size	Initial		End (at 3 month)		p-value
				mean	SD	mean	SD	
All subjects	Total IBS severity	I	45	350.41	42.91	265.75	100.62	0.310
		K	42	351.82	53.83	245.10	106.21	
	Abdominal pain	I	45	61.92	17.52	43.30	23.08	0.603
		K	42	64.56	17.64	39.96	26.18	
	Satisfaction	I	45	82.74	15.43	65.54	22.13	0.330
		K	42	76.58	16.79	57.61	23.55	
	Interference	I	45	74.41	13.97	56.22	22.60	0.327
		K	42	75.38	15.05	56.42	23.01	
	Frequency	I	45	62.89	23.22	52.22	32.11	0.358
		K	42	62.98	23.58	45.95	31.00	
	Abdominal distension	I	45	68.44	16.91	48.46	25.77	0.902
		K	42	72.32	16.26	45.17	27.88	
	QOL	I	43	52.91	18.52	40.43	23.33	**0.061**
		K	41	52.59	15.63	47.66	20.35	

1. Understanding the Structure of the Data

1.1 Recruitment

100 participants were recruited for the study, where 50 of which were assigned to group K, and 50 to the group I: one of these groups represent the active treatment group, while the other group is administered a placebo treatment (CQADS analysts do not know which label corresponds to which group).

The objective of this study is to examine the effect of the treatment against the (placebo) control group on severe IBS patients. It should be noted that there were 16 participants who were not classified as a severe IBS sufferer according to their pre-treatment total IBS severity scores. Participant ID 68, who had a severity score of 158, was discarded from the study; however, 15 patients whose baseline IBS severity scores ranging from 259.6 to 298 were kept for this study as the severity of IBS is known to fluctuate rather frequently.

1.2 Randomization

In order to facilitate a balanced representation in the active treatment group and the placebo group in terms of their demographical characteristics, participants were first categorized by their gender group (M/F) and age group (< or ≧50 years). Within each subgroup, participants were then randomly assigned to the treatment group or the placebo group, in a double-blind fashion (i.e. neither the examiners nor the participants were aware of the groups to which they had been assigned). As the number of treatment/placebo assignments in each group was not intended to be even, this randomization process leads us to (Unbalanced) Randomized Complete Block Design.

1.3 Outcome Measures

The two main response variables under considerations are the total IBS severity score and the IBS Quality of Life (QoL) measure. Furthermore, we will be examining the effect of treatment on each of the five questions that constitute the total IBS severity score. These questions measure the levels of abdominal pain, abdominal distension and bloating, satisfaction, interference, and frequency. All scores are collected at the beginning of the study (baseline) and at one-month intervals for three months. As a side note, all of these response variables are computed using self-reported data.

1.4 Drop-outs, Missing Observations, and Imputation

Eight participants did not deliver any information after the baseline measure: four participants from the group K and four from group I. As there was no information regarding the treatment effects for those participants, they were eliminated from the remaining analysis. Furthermore, six participants failed to follow-up after the first or the second month of the study. **Table 1** summarizes the breakdown of those participants.

Table 1 – IBSS drop-out data. Only those participants that remain after the first two months are retained

	Total # of recruited participants	Dropped out after Baseline	Dropped out after Month 1	Dropped out after Month 2	Remaining after Month 3
Treatment K	49	4	3	2	40 (81.6%)
Treatment I	50	4	1	0	45 (90.0%)
Total	**99**	**8**	**4**	**2**	**85 (85.8%)**

Since the covariance analysis requires the dataset to be free of missing observations, imputations must be performed prior to proceeding with the analysis.

In general, it is difficult to study the exact reasons why some participants terminate the follow-up prematurely; however it could be conjectured that participants who complete the study are either more likely to believe in the effect of the active agent or to actually be feeling the effect of the treatment than those who fail to complete the treatment. In fact, taking a look at drop-outs with partial information, it is often the case that these observations do not follow the general downward trend seen in the participants with the complete information. In an attempt to test this conjecture, partial non-respondents should be kept in the analysis.

Therefore, for those participants with recorded observations up to the second follow-up, the Last Observation Carried Forward (LOCF) imputation was favoured over the regression imputation [5], and implemented for the analysis. However, it should be noted that four participants dropped out of the study after the first follow-up. Due to the observed month-to-month fluctuation in the scores within each patient, it may not be reasonable to assume that the IBS severity scores and QoL measures for these participants stay constant over a two month period. Therefore, the decision was made to eliminate these participants from subsequent analysis.

To compensate for the fact that the imputation was done prior to the covariance analysis, one degree of freedom is docked for each imputation. Note that only the missing observations at the third month into the study are imputed, as we are interested in comparing the baseline measures and the final measures.

For the IBS severity score and its five sub-scores, there were no partial non-respondent; however, subjects 19, 22, and 32 did not complete some questions on the QoL questionnaire at the baseline. For this reason, these participants are removed from the covariance analysis for the QoL scores. **Table 2** summarizes the participants who dropped out prior to completion of the study and who were kept for the analysis with imputed scores.

Table 2 – Number of participants used in covariance analyses for IBS severity measure and QoL measure

	IBS		QoL	
Treatment group	**K**	**I**	**K**	**I**
Removed	7	5	8	7
Completed (+ imputed)	40 (42)	45	39 (41)	43
Total (Recruited)	49	50	49	50

1.5 Outlier Detection

Outlying observations frequently have a dramatic effect on the fitted values of the selected model; should such extreme points be found in the dataset, they need to be studied carefully in order to determine whether they should be retained or removed [6]. If influential observations are identified, remedial measures may need to be applied in order to minimize their undue effects.

Given that we have at most four data points per participant, and due to the large observed within-participant variability over time, it is near impossible to identify within-participant observations which we could deemed to be "extreme". It is, however, significantly easier to identify any abnormal between-participant observations.

Numerous methods can be used to find outliers; none of them are foolproof and good judgement must be used. For this reason, the box-and-whisker plots can help in the search for possible outliers: data points falling below $Q_1 - 1.5 \cdot \text{IQR}$ or above $Q_3 + 1.5 \cdot \text{IQR}$, (where Q_1, Q_3, IQR are the first quartile, the third quartile and the inter-quartile range, respectively)

require a more in-depth analysis (see **Figure 1**, on page 6). From the box-and-whisker plots, we observe that medians for treatment groups I and K usually do not differ greatly at the third follow-up. Furthermore, the variability of the data (given by the range of the whisker) tends to be greater at the last follow-up compared to the variability observed at the pre-treatment assessment.

2. Model Selection

As mentioned in **Section 1.2**, the participants were stratified according to their gender (M/F) and age group (< or ≥50 years), and then randomized within each block in an effort to promote balanced representation between two treatment groups. From a statistical perspective, blocking is used to isolate controllable variables that are not of the primary interest: since participants were randomized within each block (subgroup), and the number of treatment/placebo assignments in each group was not intended to be even, this randomization process leads us to unbalanced Randomized Complete Block Design (RCBD).

2.1 ANCOVA Models

On top of the treatment and the block effects, ANCOVA models involve the linear effect of a continuous covariate [7]: the models that we use are of the following form:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma x_{ijk} + \varepsilon_{ijk},$$

where

- y_{ijk} is the k^{th} **response variable** in the i^{th} treatment group and j^{th} block (the scores at third follow-up);
- μ is the **overall mean**;
- τ_i is the i^{th} **treatment effect**;
- β_j is the j^{th} **block effect**;
- γ is the **covariate (or regression) effect**;
- $x_{ijk} = X_{ijk} - \bar{X}$ is the k^{th} **covariate (or concomitant variable)** in the i^{th} treatment group and j^{th} block (the baseline IBSS or QoL value adjusted for the mean), and
- ε_{ijk} is the k^{th} **residual** in the i^{th} treatment group and j^{th} block.

The indices correspond to $i = 1,2,\ j = 1,\dots,4,\ k = 1,\dots,n_{ij},\ \sum_i \sum_j n_{ij} = N$, where N is the number of participants.

2.2 ANCOVA Assumptions

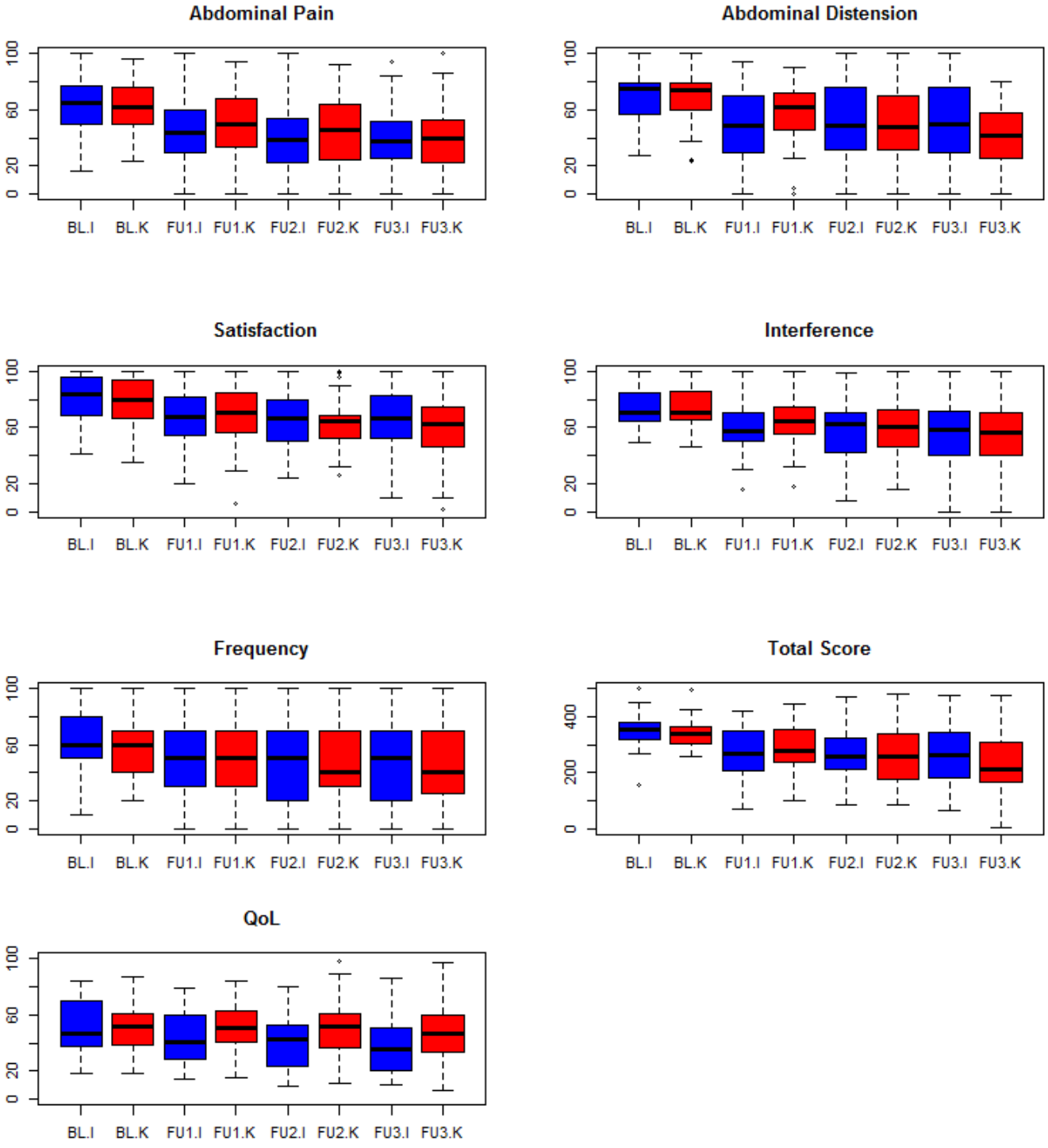
In order to use an ANCOVA model, four assumptions must be satisfied:

1. *Independence and Normality of Residuals*: the residuals are thought to be independently and identically distributed random variables following a normal distribution with zero mean (i.e. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_{\varepsilon}^2 \boldsymbol{I})$);
2. *Homogeneity of Residual Variances*: the variance of the residuals must be uniform across treatment groups;
3. *Homogeneity of Regression Slopes*: the regression effect (slope) must be uniform across treatment groups, and
4. *Linearity of Regression*: the regression relationship between the response and the covariate must be linear.

The first of these assumptions can be tested with the help of a **QQ-plot** and a scatter plot of **residual vs. fitted values**, while the second may use the **Bartlett's** or the **Levene's** test. The final assumption is not as crucial as the other three assumptions. Various remedial methods can be applied should any of these assumptions fail [6].

The third assumption, however, is critical to the ANCOVA model. It can be tested with the **equal slope test**: we run an ANCOVA regression on the models given in Sections 4 and 5 with an additional interaction term $x \times \tau$. If the interaction is not significant, the third assumption is satisfied. In the event that the interaction term is statistically significant, a different approach (e.g. moderated regression analysis, mediation analyses) is required as using the original ANCOVA model is not prescribed [8]. ANCOVA assumptions will be verified for both IBSS and QoL response variables in sections 4 and 5 respectively.

Figure 1 – Box-and-whisker plots for IBS severity scores at each time point. The blue and red columns represent the scores for treatment groups I, and K, respectively, while circles represent outlying values according to the box-and-whisker test



3. Covariance Analysis for the IBS Severity Score

A total of 100 participants were recruited for the study. One subject did not meet the recruitment criteria, and eight of which dropped out after the baseline assessment. A further three drop-outs were removed (see **Section 1.4**), leaving a total of $N = 88$ participants for the analyses for the IBS severity score and its sub scores. In order to accommodate the two imputations (again, see Section 2.4), two degrees of freedom are docked from the residual source in the ANCOVA analyses.

3.1 Total IBS Severity Score

The ANOVA table for the **ANCOVA Model on the total IBS severity score** is found in **Table 3**. At first glance, as the *p*-value for the treatment effect is 0.310, we conclude that there is not enough evidence to suggest that the two treatment effects differ at 0.05 significance level. Since the 95% confidence interval for the difference in the treatment effects include 0, the estimated treatment effects are not presented.

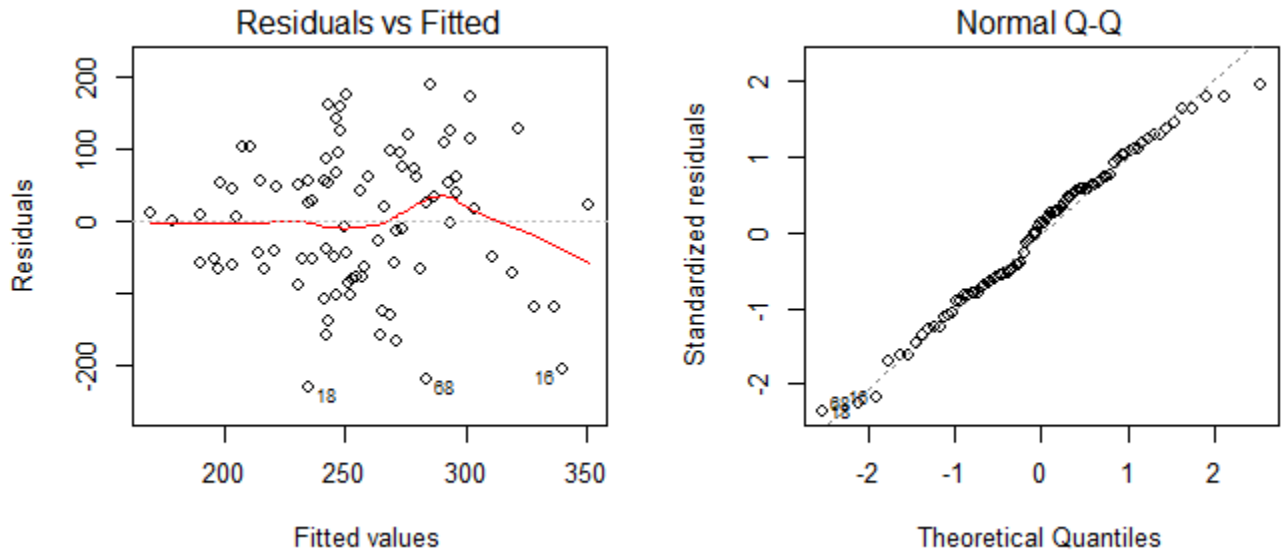
Table 3 – ANOVA table for the variance analysis on the total IBS severity score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	10521	10521	1.043	0.310
β (Block)	3	19551	6517	0.646	0.588
γ (Covariate)	1	89895	89895	8.911	0.004
ε (Residual)	81-2=79	796996	10088.56		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 2**. The data is well behaved on the normal Q-Q plot, verifying that the assumption of normality is met.

Bartlett's test is used to assess the homogeneity of the residual variances in groups K and I. The test statistic $X^2 = 0.265$, with a corresponding *p*-value of 0.60, implies that there is insufficient evidence to reject the assumption of homogeneity of variances. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 3**).

Figure 2 – Normality and independence of the residuals from ANCOVA for the total IBS severity score



Furthermore, with a *p*-value of 0.004 for the covariate effect, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear.

Finally, the test for equal slopes compares the original model $y \sim \tau + \beta + \gamma x$ to the modified interaction model

$$y \sim \tau + \beta + \gamma x + \rho(x \times \tau).$$

The lack of significance of the interaction term is interpreted as favourable to the third assumption. The appropriate ANOVA table is shown in **Table 4**; the corresponding *p*-value of 0.937 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Figure 3 – Homogeneity of variance between treatment groups I and K for the total IBS severity score based on ANCOVA

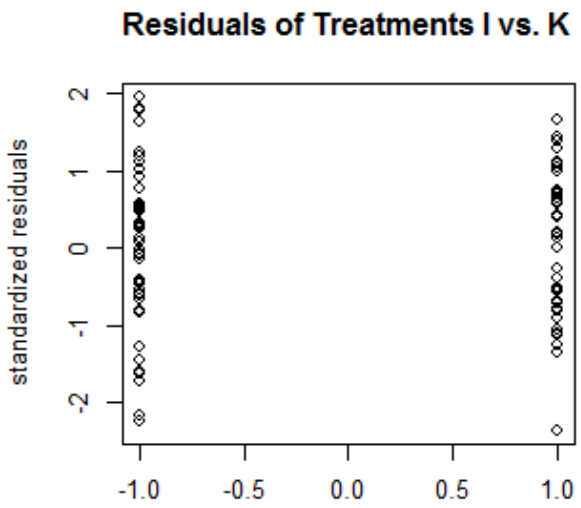


Table 4 – Homogeneity of regression slopes across treatment groups for the covariance model for the total IBS severity score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	796996				
Interaction	78	796932	1	64	0.006	0.937

The plot of residuals vs. fitted values (**Figure 2**, left) shows three outliers based on the covariance analysis. **Table 5** summarizes treatment effects on these participants. This combination provides an impetus to study the effect of possible influential observations. Note that all three outliers in **Table 5** have large reduction in the IBS severity score to categorize those participants as either not suffering from IBS (scores ranging from 0 to 75) or mildly suffering from IBS (scores ranging from 75 to 175). While their rate of reduction is anomalous compared to the rest of the participants, since not all three participants belong to one group, the covariance analysis on reduced dataset (i.e., IDs 16, 18, and 68 removed) should not alter the results significantly. Hence, no further analyses are conducted for the total IBS severity score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 5 – Outliers based on the covariance analysis on the total IBS severity score

ID	Group	Baseline score	Final score	Difference
16	I	448	134	-314
18	K	326	6	-320
68	I	365	65	-300

3.2 Abdominal Pain Score

The ANOVA table for the **abdominal pain score using ANCOVA Model** is found in **Table 6**. It should be noted that the *p*-value for the covariate effect is 0.630, the result suggests that analysis of variance would be more appropriate than analysis of covariance to test the difference in the abdominal pain scores in two treatment groups.

Table 6 – ANOVA table for the covariance analysis on the abdominal pain score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	192	192	0.314	0.577
β (Block)	3	3003	1001	1.639	0.187
γ (Covariate)	1	143	143	0.233	0.630
ε (Residual)	81-2=79	48261	611		

Table 7, which provides the ANOVA table for the analysis of variance on the abdominal pain score, indicates that the treatment effects do not differ as the *p*-value for the difference in the treatment effects is 0.603. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 4**. The normal Q-Q plot shows a slight deviation from the assumption of normality; however, as ANOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Table 7 – ANOVA table for the variance analysis on abdominal pain scores with degrees of freedom modified to accommodate imputation.

Source	**df**	**Type III SS**	**MS**	***F***	***p*-value**
$\boldsymbol{\tau}$ (Treatment)	1	163.31	163.31	0.273	0.603
$\boldsymbol{\beta}$ (Block)	3	3147.44	1049.15	1.759	0.162
$\boldsymbol{\varepsilon}$ (Residual)	83-2=81	48404	597.58		

To assess the homogeneous variances of the residuals in the groups I and K, Bartlett's test is used. There is insufficient evidence to conclude that the variances are non-homogeneous across treatment groups as the statistic is $X^2 = 0.239$ with a corresponding *p*-value of 0.625. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 5**).

Figure 4 – Normality and independence of the residuals from ANOVA for the abdominal pain scores

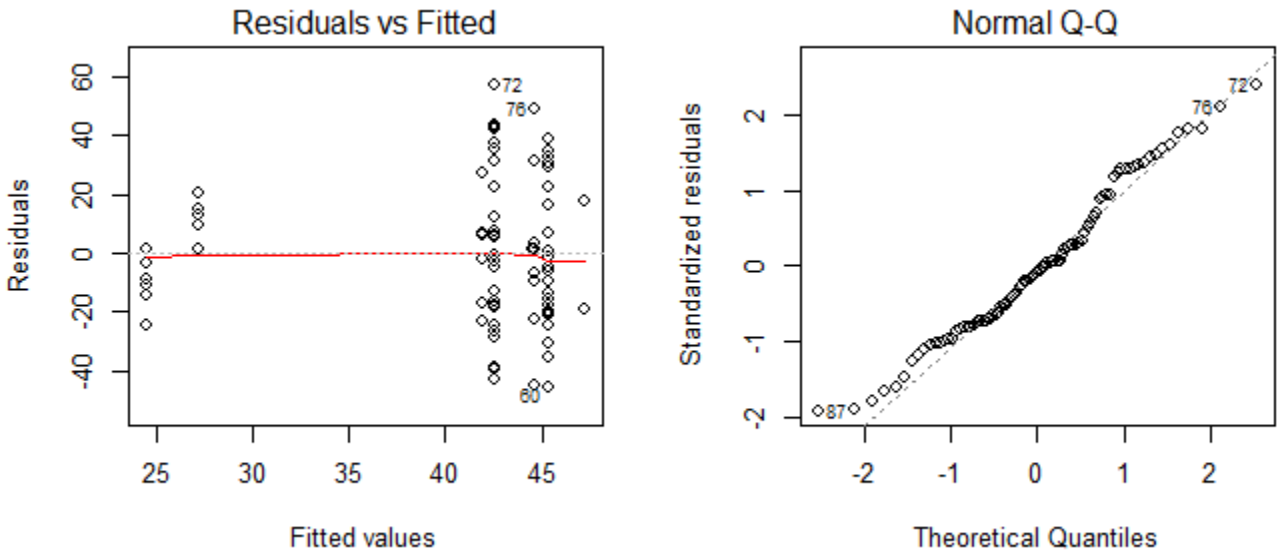
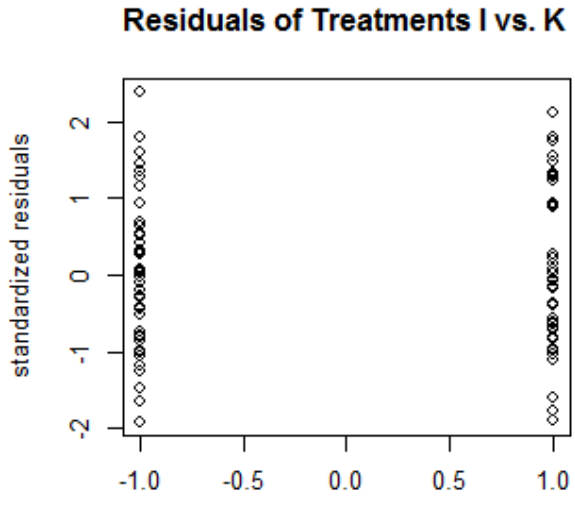


Figure 5 – Homogeneity of variance between treatment groups I and K for the ANOVA of the abdominal pain score



The plot of residuals vs. fitted values (**Figure 4**, left) shows three outliers based on the variance analysis. **Table 8** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the *p*-value associated with the treatment effect on the abdominal pain score is 0.603, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 8 – Outliers based on the analysis of variance on the abdominal pain score

ID	Group	Baseline score	Final score	Difference
73	K	50	100	50
77	I	78	94	16
88	I	76	0	-76

3.3 Satisfaction Score

Table 9 provides the ANOVA table for the **satisfaction score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.330, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 6**. The normal Q-Q plot demonstrates deviation from the assumption of normality on both tails; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a moderate deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.072$ with a corresponding *p*-value of 0.790. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 7**).

Table 9 – ANOVA table for the covariance analysis on satisfaction score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	837	837	0.961	0.330
β (Block)	3	4089	1363	1.565	0.205
γ (Covariate)	1	13078	13078	15.013	<0.001
ε (Residual)	81-2=79	68815	871		

Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 10**; the corresponding *p*-value of 0.261 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 6**, left) shows three outliers based on the covariance analysis. **Table 11** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.330, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 6 – Normality and independence of the residuals from ANCOVA for the satisfaction score

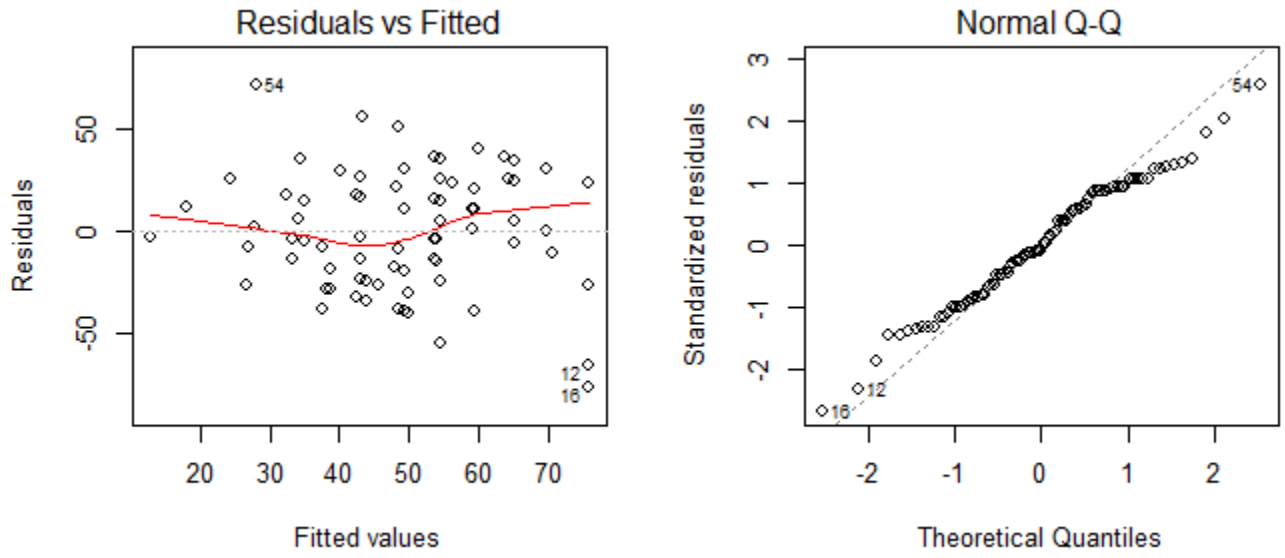


Figure 7 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the satisfaction score

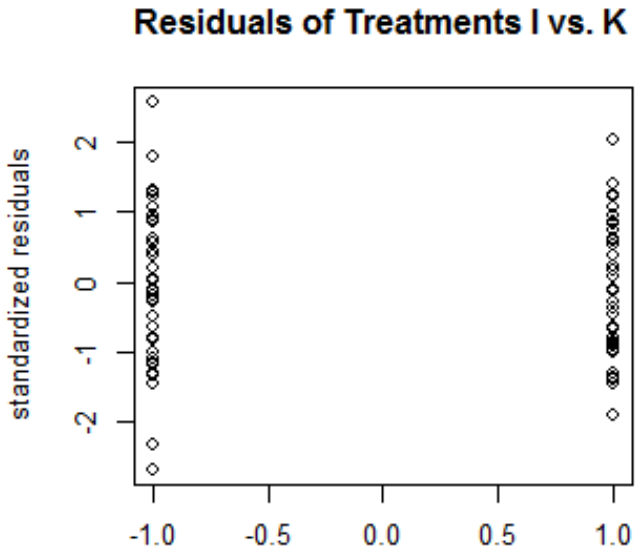


Table 10 – Homogeneity of regression slopes across treatment groups for the covariance model for the satisfaction score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	68815				
Interaction	78	67700	1	1115	1.280	0.261

Table 11 – Outliers based on the analysis of variance on the satisfaction score

ID	Group	Baseline score	Final score	Difference
12	I	100	10	-90
16	K	100	0	-100
55	I	10	100	90

3.4 Interference Score

Table 12 provides the ANOVA table for the **interference score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.327, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

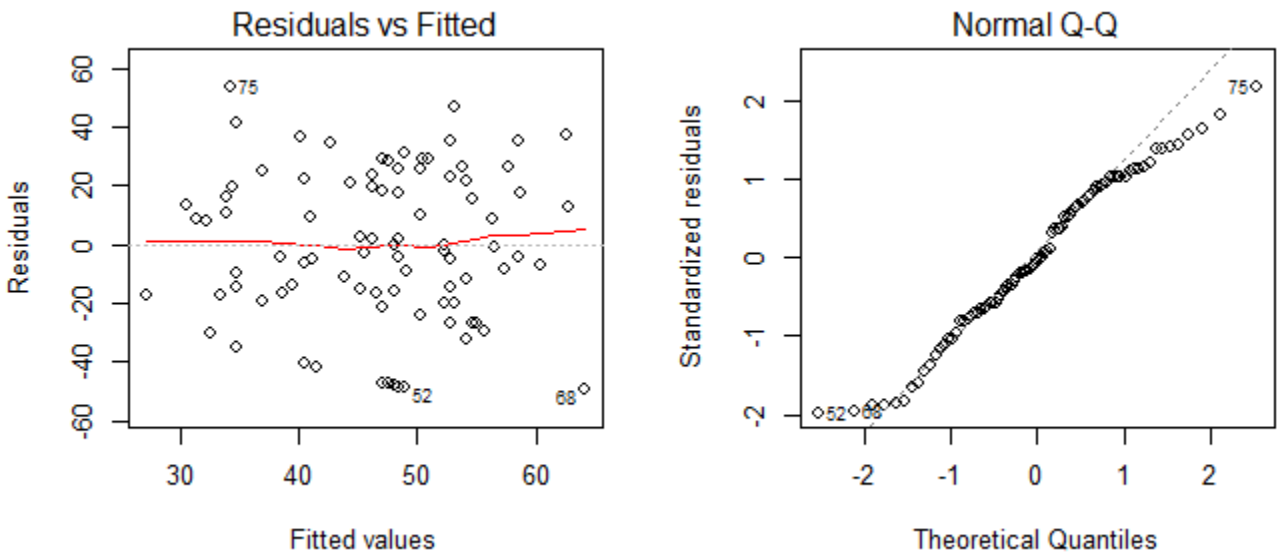
Table 12 – ANOVA table for the covariance analysis on interference score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	680	680	0.973	0.327
β (Block)	3	878	293	0.419	0.740
γ (Covariate)	1	4899	4899	7.013	0.010
ε (Residual)	81-2=79	55183	699		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 8**. The normal Q-Q plot demonstrates deviation from the assumption of normality on both tails; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a moderate deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.068$ with a corresponding *p*-value of 0.795. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 9**).

Figure 8 – Normality and independence of the residuals from ANCOVA for the interference score



Furthermore, with a *p*-value for the covariate effect being 0.01, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 13**; the corresponding *p*-value of 0.261 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 8**, left) shows three outliers based on the covariance analysis. **Table 14** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the interference score is 0.327, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 9 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the interference score

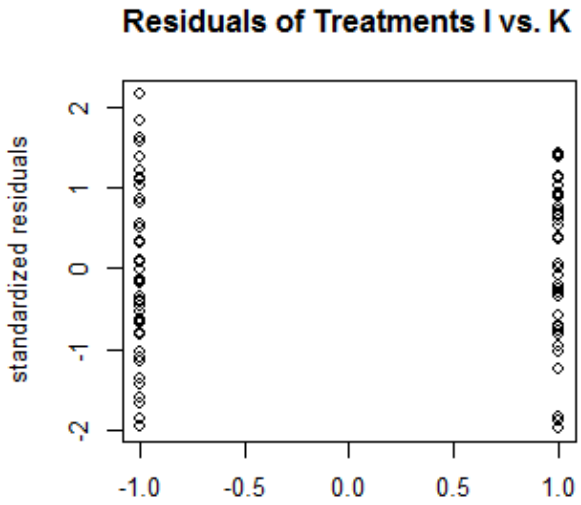


Table 13 – Homogeneity of regression slopes across treatment groups for the covariance model for the interference score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	68815				
Interference	78	67700	1	1115	1.280	0.261

Table 14 – Outliers based on the analysis of variance on the interference score

ID	Group	Baseline score	Final score	Difference
53	K	87	0	-87
69	I	100	15	-85
76	K	56	88	32

3.5 Frequency Score

Table 15 provides the ANOVA table for the **frequency score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.358, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

Table 15 – ANOVA table for the covariance analysis on frequency score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
$\boldsymbol{\tau}$ (Treatment)	1	596	596	0.854	0.358
$\boldsymbol{\beta}$ (Block)	3	1116	372	0.533	0.661
$\boldsymbol{\gamma}$ (Covariate)	1	3588	3588	7.083	0.009
$\boldsymbol{\varepsilon}$ (Residual)	81-2=79	40014	507		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 10**. The normal Q-Q plot demonstrates a slight deviation from the assumption of normality; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a minor deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.321$ with a corresponding *p*-value of 0.573. There is thus

insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 11**).

Figure 10 – Normality and independence of the residuals from ANCOVA for the frequency score

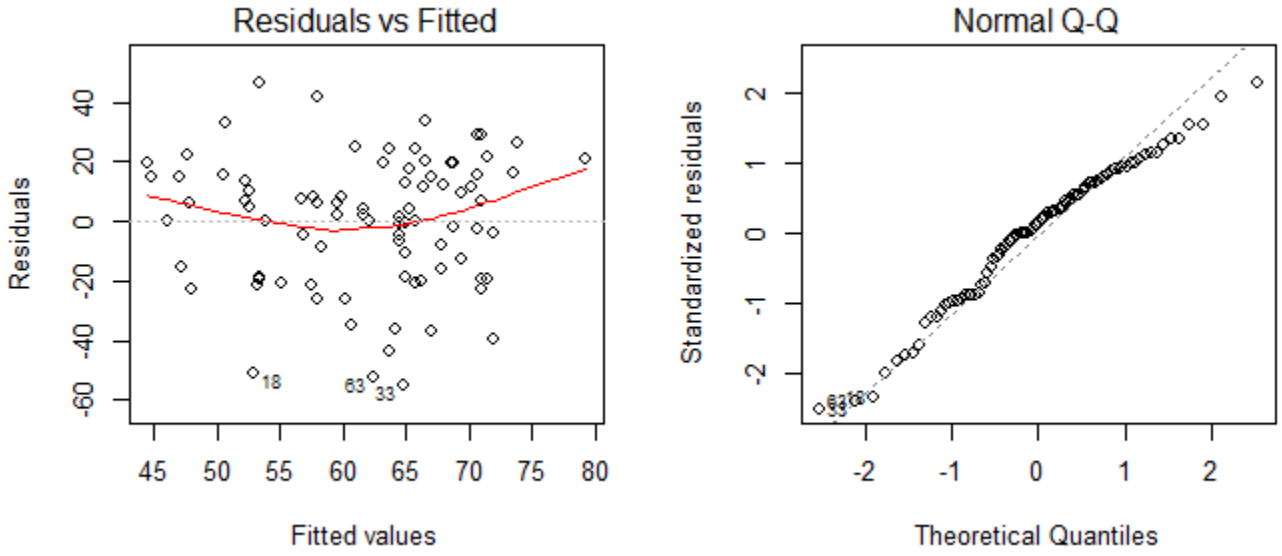
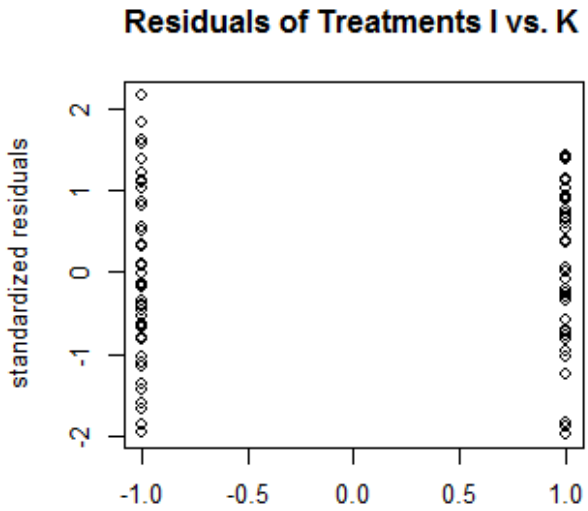


Figure 11 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the frequency score



Furthermore, with a *p*-value for the covariate effect being 0.009, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 16**; the corresponding *p*-value of 0.427 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 10**, left) shows three outliers based on the covariance analysis. **Table 17** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.358, analysis on the reduced dataset (i.e., potential influential observations removed) should not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Table 16 – Homogeneity of regression slopes across treatment groups for the covariance model for the frequency score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	40014				
Frequency	78	40006	1	8.000	0.016	0.427

Table 17 – Outliers based on the analysis of variance on the frequency score

ID	Group	Baseline score	Final score	Difference
18	K	66.7	2	-64.7
34	I	82.7	10	-72.7
64	K	90	10	-80

3.6 Abdominal Distension Score

Table 18 provides the ANOVA table for the **abdominal distension score using ANCOVA Model**. As the *p*-value for the treatment effect is given to be 0.902, we conclude that there is not enough evidence to suggest that the treatment has an effect at the 0.05 significance level.

Table 18 – ANOVA table for the covariance analysis on abdominal distension score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	7	7	0.015	0.902
β (Block)	3	847	282	0.586	0.626
γ (Covariate)	1	5383	5383	11.182	0.001
ε (Residual)	81-2=79	38028	481		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 12**. The normal Q-Q plot demonstrates a slight deviation from the assumption of normality; however, as ANCOVA is moderately robust to the violation of this assumption, the level of deviation seen here is no concern.

Due to a minor deviation from the normality assumption, Levene's test is used to assess the homogeneous variances of the residuals in the groups K vs. I. The test statistic is $W = 0.059$ with a corresponding *p*-value of 0.809. There is thus insufficient evidence to conclude that the variances are non-homogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 13**).

Furthermore, with a *p*-value for the covariate effect being 0.001, it seems reasonable to assume that the relationship between the response and the covariate is linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 19**; the corresponding *p*-value of 0.835 indicates that that it is reasonable to assume the homogeneity of regression slopes.

The plot of residuals vs. fitted values (**Figure 12**, left) shows three outliers based on the covariance analysis. **Table 20** summarizes treatment effects on them. This combination provides an impetus to study the effect of possible influential observations. However, since the p-value associated with the treatment effect on the satisfaction score is 0.358, analysis on the reduced dataset (i.e., potential influential observations removed) would not result in change in the decision based on ANOVA. Therefore, no further analyses are conducted for the abdominal pain score and we conclude that there is not enough evidence to believe that treatments I and K produces significantly different results.

Figure 12 – Normality and independence of the residuals from ANCOVA for the abdominal distension score

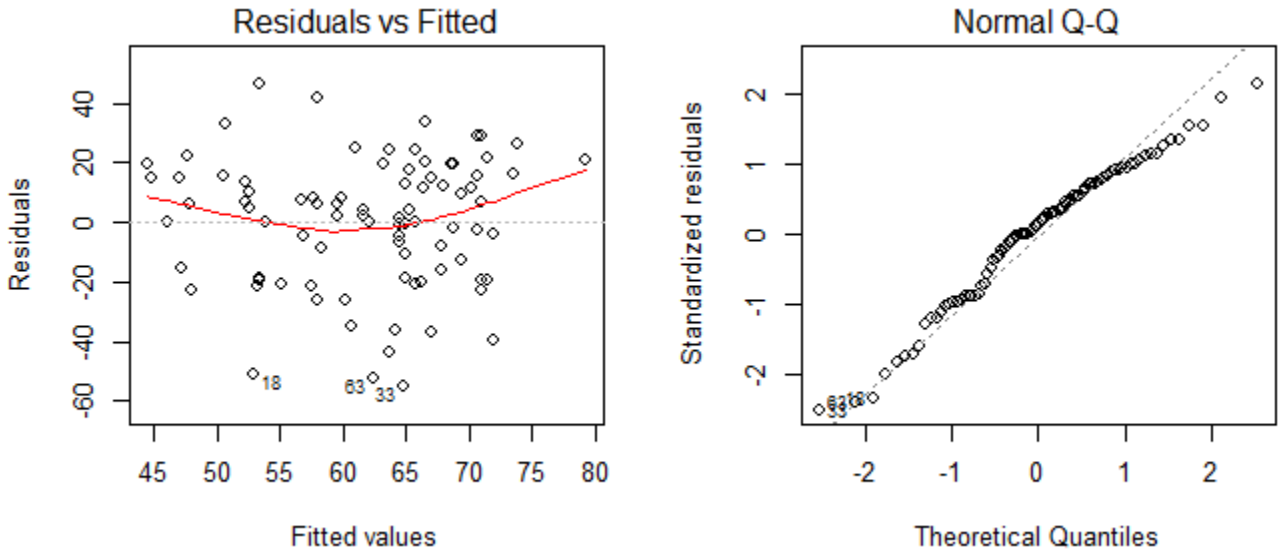


Figure 13 – Homogeneity of variance between treatment groups I and K for the ANCOVA of the abdominal distension score

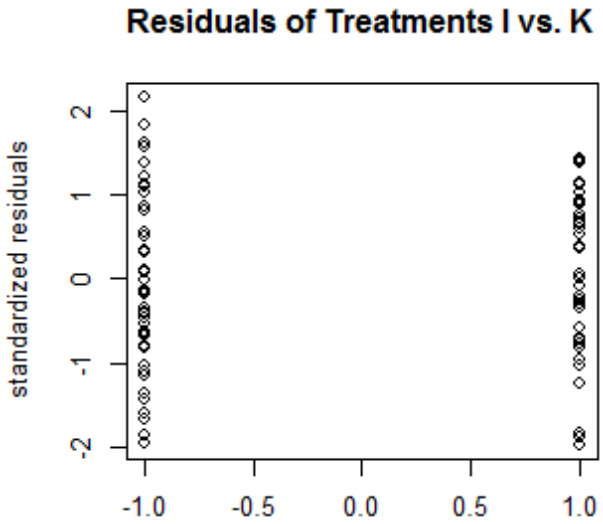


Table 19 – Homogeneity of regression slopes across treatment groups for the covariance model for the abdominal distension score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	79	38028				
Interaction	78	38007	1	21.000	0.044	0.835

Table 20 – Outliers based on the analysis of variance on the abdominal distension score

ID	Group	Baseline score	Final score	Difference
18	K	66.7	0	-66.7
64	I	80	10	-70
69	I	75	5	-70

4. Covariance Analysis for the QoL Score

As before, a total of 100 participants were recruited for the study, where one subject did not meet the recruitment criteria, three subjects had incomplete baseline measure for QoL, and eight of which dropped out after the baseline assessment. A further four drop-outs were removed (see Section 2.4), leaving a total of $N = 84$ participants for the analyses for the IBS severity score and its sub scores. In order to accommodate the two imputations (again, see Section 2.4), two degrees of freedom are docked from the residual source in the ANCOVA analyses.

4.1 QoL Score on Full Dataset

The ANOVA table for the **ANCOVA Model on the QoL score** is found in **Table 21**. At first glance, as the *p*-value for the treatment effect is 0.061, we conclude that there is not enough evidence to suggest that the two treatment effects differ at 0.05 significance level; however, it should be noted that the point estimate yields that, on average, participants in treatment group I have lost an extra 7.26 QoL score over the course of three months treatment period.

Table 21 – ANOVA table for the variance analysis on the QoL score with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	1099	1099	3.629	0.061
β (Block)	3	370	123	0.407	0.748
γ (Covariate)	1	14847	14847	49.031	<0.001
ε (Residual)	78-2=76	23013	303		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 14**. The data is well behaved on the normal Q-Q plot, verifying that the assumption of normality is met.

Bartlett's test is used to assess the homogeneous variances of the residuals in the groups K vs. I. The test statistic is $X^2 = 0.006$, with a corresponding *p*-value of 0.937 imply that there is insufficient evidence to conclude that the variances are heterogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 15**).

Figure 14 – Normality and independence of the residuals from ANCOVA for the QoL score

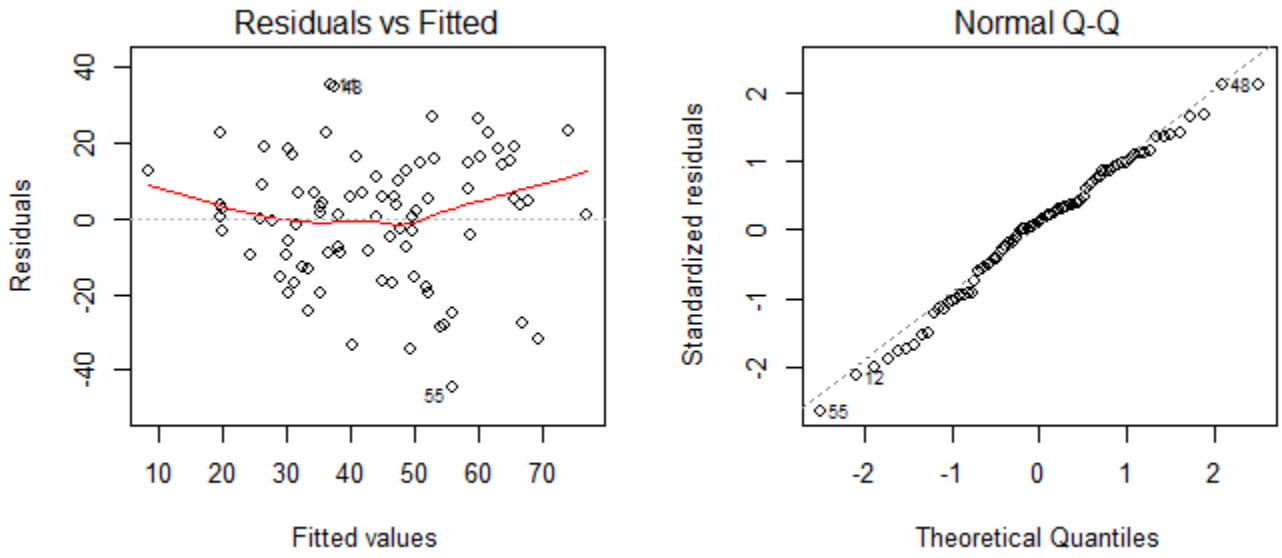
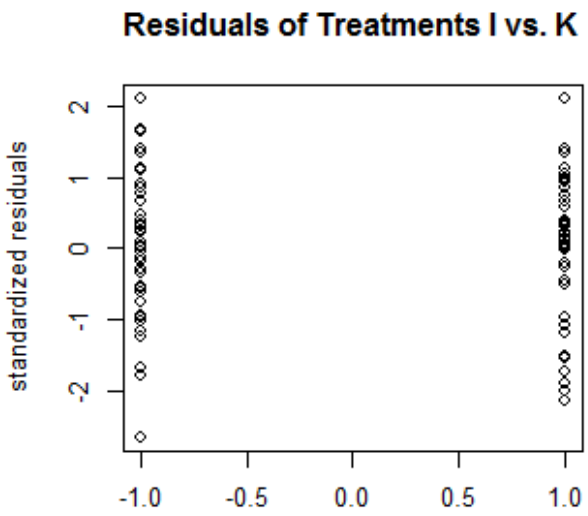


Figure 15 – Homogeneity of variance between treatment groups I and K for the QoL score based on ANCOVA



Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 22**; the corresponding *p*-value of 0.481 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Table 22 – Homogeneity of regression slopes across treatment groups for the covariance model for the QoL score with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	76	23014				
Interaction	75	22862	1	152.000	0.502	0.481

The plot of residuals vs. fitted values (**Figure 14**, left) shows three outliers based on the covariance analysis. **Table 23** summarizes treatment effects on these participants. This combination provides an impetus to study the effect of possible influential observations. Since the p-value associated with the difference in the effects of the two treatment groups is 0.061, we will examine whether the treatment effect would be statistically significant, under the removal of the potential influential observations.

Table 23 – Outliers based on the covariance analysis on the QoL score

ID	Group	Baseline score	Final score	Difference
14	K	37.5	72.1	34.6
59	I	54.4	72.1	11.7
69	I	70.2	11.4	-58.8

4.2 QoL Score on Reduced Dataset

The ANOVA table for the **ANCOVA Model on the QoL score based on a reduced dataset** is found in **Table 24**. At first glance, as the *p*-value for the treatment effect is increased to 0.093, we conclude that there is not enough evidence to suggest that the two treatment effects differ at a 0.05 significance level; however, it should be noted that the point estimate yields that, on average, participants in treatment group I have lost an extra 6.04 QoL score over the course of three months treatment period.

Table 24 – ANOVA table for the variance analysis on the QoL score on a reduced dataset with degrees of freedom modified to accommodate imputation.

Source	df	Type III SS	MS	*F*	*p*-value
τ (Treatment)	1	733	733	2.906	0.093
β (Block)	3	730	243	0.965	0.414
γ (Covariate)	1	16494	16494	65.381	<0.001
ε (Residual)	75-2=73	18416	252		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 16**. The data shows a minor deviation from the assumption of normality on the normal Q-Q plot; however, as the ANCOVA is moderately robust to the deviation from the normality assumption, the level of deviation seen here is no concern.

The Levene's test is thus used to assess the homogeneous variances of the residuals in the groups I and K. The test statistic is $W = 0.023$, with a corresponding p-value of 0.881, implying that there is insufficient evidence to conclude that the variances are heterogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 17**).

Figure 16 – Normality and independence of the residuals from ANCOVA for the QoL score

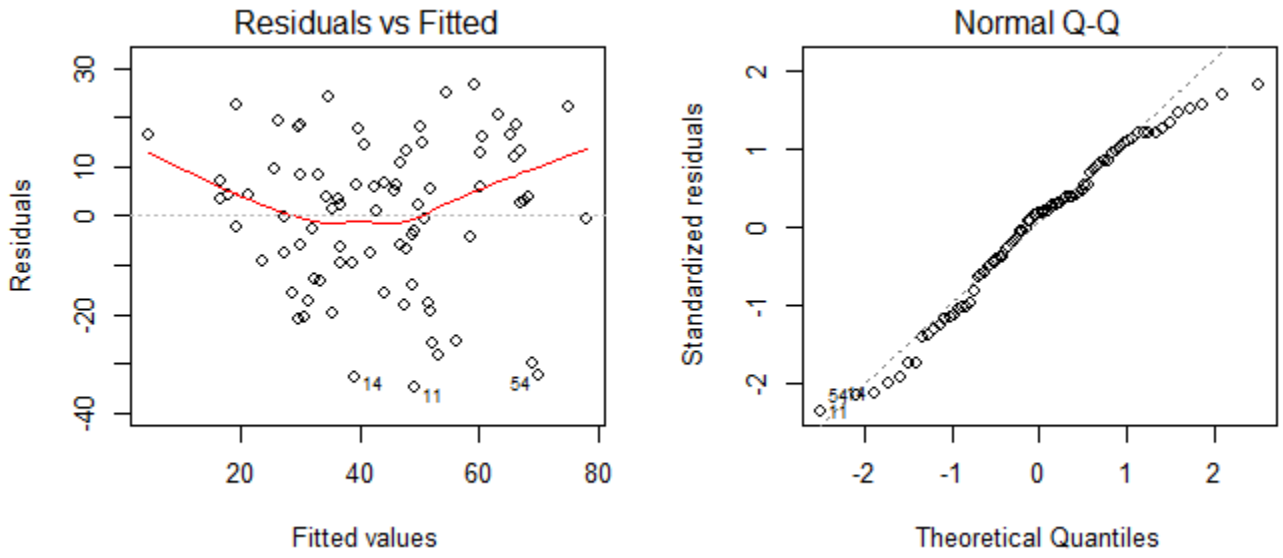
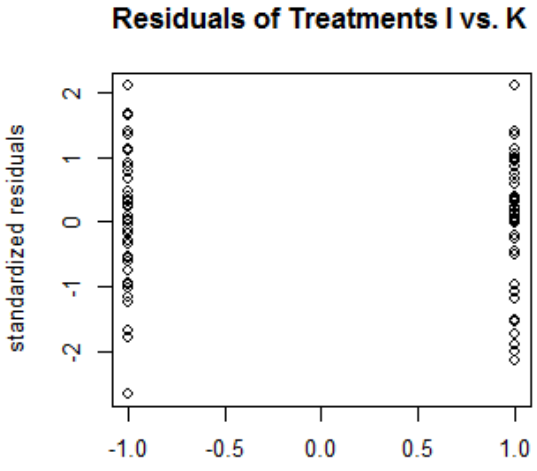


Figure 17 – Homogeneity of variance between treatment groups I and K for the QoL score based on ANCOVA



Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear. The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 25**; the corresponding *p*-value of 0.467 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Table 25 – Homogeneity of regression slopes across treatment groups for the covariance model for the QoL score on a reduced dataset with degrees of freedom modified to accommodate imputation.

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	73	18416				
Interaction	72	18281	1	135.000	0.535	0.467

The plot of residuals vs. fitted values (**Figure 16**, left) shows three outliers based on the covariance analysis. However, since these observations are classified as an outlier only due to the removal of the three outliers from the original dataset, we will not perform any further analyses on the QoL score, and conclude that, at 95% significance level, two treatment groups do not differ in their treatment effects.

5. IBS Sub-Score Analyses for 2010 Dataset

Extremely similar analyses were conducted for the sub-scores of the IBS data collected during the 2010 pilot study; in the interest of readability, the results were condensed and placed in a table format in the Executive Summary. While none of the sub-scores showed statistically significant improvement under the probiotic agent, one of them (Statisfaction, *p*-value: 0.085) was nearly significant.

6. Conclusions and Recommendations

We end the report with key findings of our analysis, as well as some recommendations for future investigations.

6.1 Blocking and Balanced Designs

In this report, we have found that blocking (or subgrouping) the participants according to their gender and age does not play an important role in the ANCOVA. In future studies involving this probiotic agent, blocking should only be used if there are compelling reasons to suspect that treatment effects are different for at least one subgroup, as blocking results in fewer degrees of freedom.

Special care should also be taken to have a balanced design (i.e., equal number of replicates for each subgroup), especially if subgroup analyses are of interest: for instance, the overwhelming number of female participants and small number of male participants make any conclusions about male subgroups statistically unsound.

6.2 Recruitment Process

In the 2013 IBS Study, participants needed to come forward to be selected. The recruitment process used advertisements on the radio, in local newsletters and newspapers, on the web and social media, as well as posters with which local MDs and NDs could encourage patient referrals.

The elephant in the room is that this type of recruitment process leads to self-selection biases: the participants in the 2013 IBS Study may not constitute a representative sample of IBS sufferers, which makes it difficult to generalize the result of the analyses beyond the collected sample, even when there is a significant impact.

This is a problem that plagues numerous clinical studies – unfortunately, it is quite difficult to counter this situation.

6.3 Practical Significance of Results

With the caveat brought up in section 6.2, our interpretation of the covariance analyses results is that there is simply not enough evidence to conclude that the agent is effective against IBS.

It is true that the difference in the treatment effects between the two groups on the (self-reported) QoL score is nearly statistically significant at the 0.05 significance level. The corresponding estimated difference in the treatment effects is 7.26 under the using full dataset, which means that on average, participants in the group I seem to have lost an extra 7.26 QoL points over the course of three months, compared to those in the group K. However, given the amount of variability in individuals from month to month, we are reluctant to conclude that the agent under investigation provides a practically significant improvement in the average participant's quality of life.

Further investigation may shed some light on the situation and will help us determine if the relationship between the agent and QoL is causal or spurious.

6.4 Publication of Results

Even though this study did not find any statistically significant improvement for IBS, it should be published in order to counter publication bias.

References

[1] "Irritable Bowel Syndrome," [Online]. Available: http://en.wikipedia.org/wiki/Irritable_bowel_syndrome. [Accessed 5 May 2013].

[2] P. Paré, J. Gray, S. Lam, R. Balshaw, S. Khorasheh, M. Barbeau, S. Kelly and C. R. McBurney, "Health-related quality of life, work productivity, and health care resource utilization of subjects with irritable bowel syndrome: baseline results from LOGIC (Longitudinal Outcomes Study of Gastrointestinal Symptoms in Canada), a naturalistic study," Clinical Therapeutics, vol. 28, no. 10, pp. 1726-35, 2006.

[3] S. Maxion-Bergemann, F. Thielecke, F. Abel and R. Bergemann, "Costs of irritable bowel syndrome in the UK and US," PharmacoEconomics, vol. 24, no. 1, pp. 21-37, 2006.

[4] P. Herman, C. Kooley and D. Seely, "Double-blind placebo-controlled pilot study to investigate the effects of an investigational Probiotic on Irritable Bowel Syndrome," 2011.

[5] S. Hagiwara, "Nonresponse Error in Survey Sampling: Comparison of Different Imputation Methods," Ottawa, 2012.

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li, Applied Linear Statistical Models, 5th ed., New York: McGraw-Hill/Irwin, 2004.

[7] P. W. John, Statistical Design and Analysis of Experiments, New York: Macmillan, 1971.

[8] S. Green and N. Salkind, Using SPSS for Windows and Macintosh: Analyzing and Understanding Data, 6th ed., Upper Saddle River, NJ: Prentice Hall, 2011.

[9] Hagiwara, S. and Boily, P., "Covariance Analysis for the 2010 CCNM Pilot Study on Irritable Bowel Syndrome," Internal Report to the CCNM (available from CQADS), 2013.

Appendix – Analyses for Participants With Severe Baseline IBS

The client has expressed interest in analyzing the data restricted to participants with severe baseline IBS (300 or higher). We've elected to first run the analysis on the most promising variable: QoL (p-value: 0.061). It should be noted that reducing the number of observations has the tendency to suppress the test statistic which make it more difficult to find significant effects.

A.1 QoL Score on Severe Baseline Dataset

As before, a total of 100 participants were recruited for the study, where 16 subjects did not meet the recruitment criteria, and eight of which dropped out after the baseline assessment. A further four drop-outs were removed (see Section 2.4), leaving a total of $N = 72$ participants for the QoL analysis.

The ANOVA table for the **ANCOVA Model on the QoL score for the participants with baseline IBS Severity score of 300 or higher** is found in **Table 26**. At first glance, as the p-value for the treatment effect is 0.242, we conclude that there is not enough evidence to suggest that the two treatment effects differ at a 0.05 significance level.

Table 26 – ANOVA table for the variance analysis on the QoL score.

Source	df	Type III SS	MS	F	p-value
τ (Treatment)	1	461.8	461.8	1.393	0.242
β (Block)	3	300.3	100.1	0.302	0.824
γ (Covariate)	1	11072.5	11072.5	30.375	<0.001
ε (Residual)	66	20568.6	311.6		

The ANCOVA assumptions are verified as follows. The assumption of independence of the residuals is satisfied based on the visual assessment of diagnostic plots in **Figure 18**. The data is well behaved on the normal Q-Q plot, verifying that the assumption of normality is met.

Bartlett's test is used to assess the homogeneous variances of the residuals in the groups K vs. I. The test statistic is $X^2 = 0.015$, with a corresponding p-value of 0.903 imply that there is insufficient evidence to conclude that the variances are heterogeneous across treatment groups. A plot of the variances corroborates the assertion that the second assumption is met (see **Figure 18**).

Figure 18– Normality and independence of the residuals from ANCOVA for the QoL score

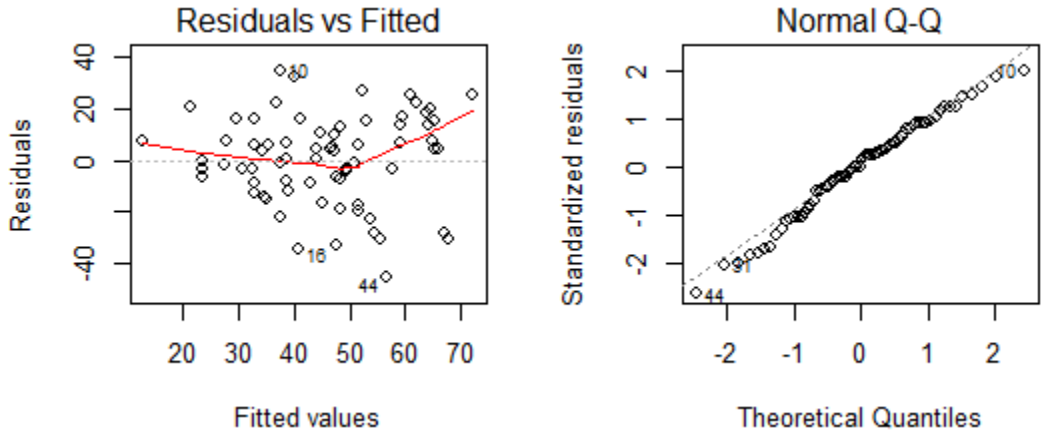
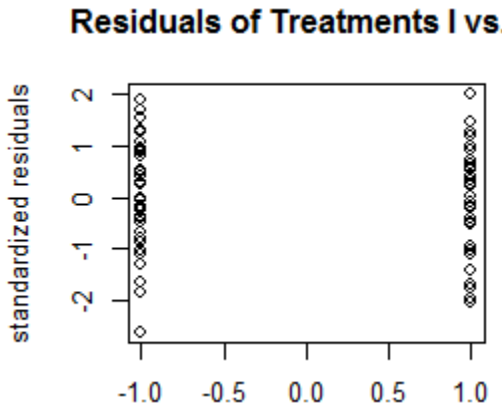


Figure 19 – Homogeneity of variance between treatment groups I and K for the QoL score based on ANCOVA



Furthermore, with a *p*-value for the covariate effect being less than 0.001, it seems reasonable to assume that the relationship between the response and the covariate is indeed linear.

The ANOVA table for the test of homogeneity of the regression slopes is shown in **Table 27**; the corresponding *p*-value of 0.6684 indicates that that it is reasonable to assume the homogeneity of regression slopes.

Table 27 – Homogeneity of regression slopes across treatment groups for the covariance model for the QoL score

Model	df_ε	RSS	df_{diff}	SS	*F*	*p*-value
Original	66	20569				
Interaction	65	20510	1	58.436	0.1852	0.6684

The plot of residuals vs. fitted values (**Figure 18**, left) shows three outliers based on the covariance analysis. **Table 28** summarizes treatment effects on these participants. This combination provides an impetus to study the effect of possible influential observations. The analysis was also run with the outliers removed: the *p*-value was even higher, reaching 0.3.

Table 28 – Outliers based on the covariance analysis on the QoL score

ID	Group	Baseline score	Final score	Difference
14	K	37.5	72.1	34.6
18	K	41.9	6.6	-35.3
69	I	70.2	11.4	-58.8

It is our contention that a similar shift for the worst will be experienced with all analyses: dropping the non-severe IBS sufferers translates to a sizeable reduction in the number of degrees of freedom, which in turn makes it much more difficult to detect a significant effect, should one even exist.