Clinical Trial Data Analysis

Kellie Ottoboni

2017-10-02

Summary

We reanalyzed a dataset from a clinical trial comparing the effectiveness of two treatments for gastroesophageal reflux disease (GERD). Patients were treated at eight sites in two different countries. At each site, patients were randomly assigned one of two treatments. Patients were observed for a week before receiving treatment and for a week after receiving treatment. On each of the fourteen days of observation, patients responded to a survey about their heartburn, regurgitation, and dyspepsia frequency and severity. These endpoints were measured on a discrete scale. There were several additional endpoints, such as daily regurgitation, daily "hrdq", and daily dyspepsia, calculated from the survey measures. Daily "hrdq" was the primary endpoint. To reduce day-to-day variation, we averaged the measures from each week to obtain two observations per patient, one pre-treatment and one post-treatment. These averages are no longer discrete ordinal values, but they are not continuous either: they can only take a finite set of values.

In what follows, we compare the parametric ANCOVA, the stratified permutation test, and the two linear model-based permutation tests. We begin by walking through the tests in detail using the primary endpoint. We illustrate different ways of controlling for the baseline measure: including it in the model and using change scores as the dependent variable. The original analysis of the dataset controls for baseline by including it in the model, so for consistency we choose that method for our final analysis. We conclude by running the tests for each endpoint.

We show that the data may violate the assumptions for the parametric ANCOVA. In particular, the outcome variables and the linear regression residuals are skewed to the right, and there is evidence of heteroskedasticity. The small p-values that come from the parametric analysis may therefore be unreliable. However, the permutation tests yield only slightly larger p-values. The correspondence between parametric and permutation test results gives some confidence that the results are stable and reliable.

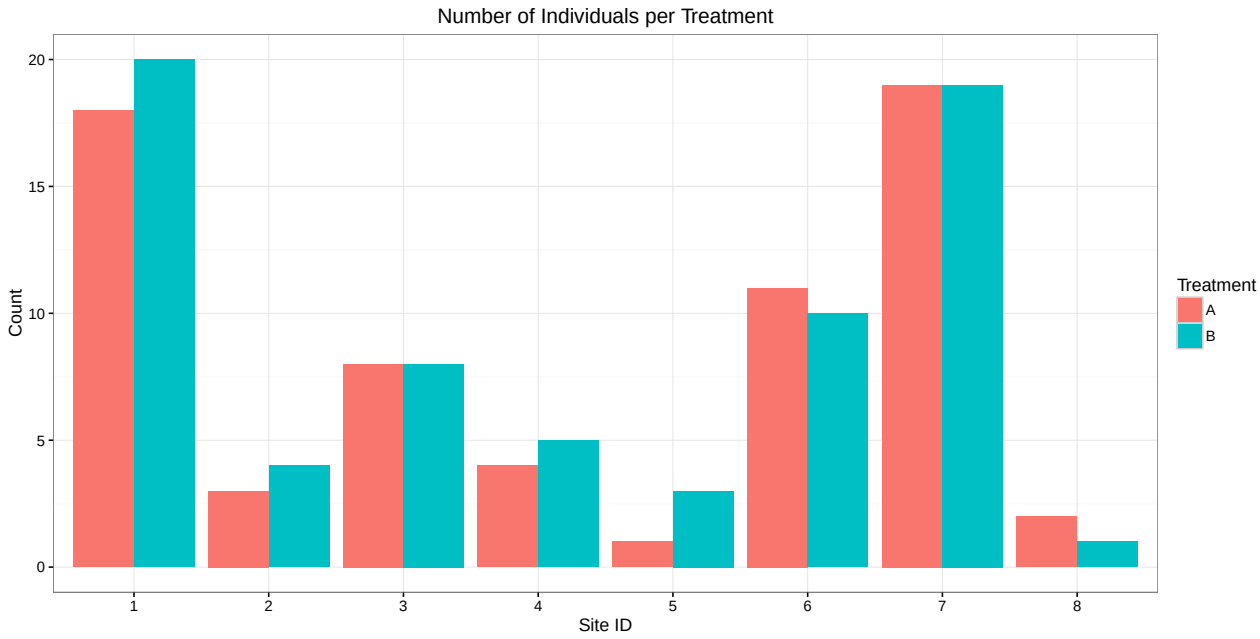

Figure 1: Number of subjects in each treatment group, by site ID.

Data Cleaning

We preprocessed the data from its original format in the file "data/clinical_cleaned.R" so that it would be sufficiently anonymized. We took the following steps:

- We renamed treatments, subject IDs, and site IDs to remove identifying information.
- Initially, there were 188 subjects. 52 of these were never assigned to a treatment group and were only observed at baseline. Because censoring occurred before treatment assignment, we are guaranteed that censoring and treatment are independent. Removing these individuals from the analysis does not introduce any selection bias. After removing the 52 people with missing data, we are left with 66 subjects in group A and 70 subjects in group B.
- Each individual was observed for a total of 14 days, 7 during the baseline week and 7 after receiving treatment. We averaged the 7 measurements taken during each of these time periods, leaving us with 2 measurements per individual.

Exploratory Data Analysis

We notice several things in Figure 1. Some of the sites have many more subjects than others: these are sites 1, 3, and 6, and 7. Sites 2, 5, and 8 have very few subjects, so we will have low power at these sites. For example, after removing the missing values, site 8 only has one person in group B and two in group A, so the site will only have 3 unique permutation statistics.

```
clinical %>% filter(VISITNUM == 1) %>% ggplot(aes(x = factor(SITEID))) +
    geom_bar(aes(fill = tr), position = "dodge") + xlab("Site ID") +
    ylab("Count") + ggtitle("Number of Individuals per Treatment") +
    labs(fill = "Treatment") + theme_bw()
```

The dataset is not in the right format to analyze a single variable. Below, we include a function to take the raw data and reshape it to analyze one variable of interest.

```r
data_by_subjid_visitnum <- clinical %>% group_by(SUBJID, VISITNUM)

reshape_data <- function(variable, data = data_by_subjid_visitnum) {
    # Reshape data to be analyzed with regression Inputs:
    # variable = the clinical endpoint of interest, input as a
    # string data = dataset with the variable as a column name,
    # grouped by subject id and visit number default is the
    # dataframe we just created, data_by_subjid_visitnum Output:
    # A dataframe with a single row per subject and columns for
    # treatment, site ID, baseline + outcome measures
    data <- data %>% mutate_(myvariable = variable)
    cleaned <- dcast(data, SUBJID + tr + SITEID ~ VISITNUM, value.var = "myvariable")
    colnames(cleaned) <- c("SUBJID", "tr", "SITEID", "Baseline",
        "Outcome")
    cleaned <- ungroup(cleaned) %>% mutate(difference = Outcome -
        Baseline)
    return(cleaned)
}
```

Figure 2 shows the distribution of the final outcome measures between the two treatment groups. Table 1 shows the means of the distributions. Based on these summary statistics, we would expect to detect a difference in outcomes between the treatments for the `daily_heart`, `daily_hrdq`, and `heart_freq` measures.

```r
# Create a plot for each variable, store in a list
continuous_vars <- c("daily_heart", "daily_regurg", "daily_dysp",
    "daily_hrdq", "heart_freq", "regurg_freq", "dysp_freq")
plot_distrs <- lapply(continuous_vars, function(variable) {
    p <- reshape_data(variable) %>% mutate(tr = factor(tr)) %>%
        ggplot(aes(Outcome)) + geom_density(alpha = 0.6, aes(fill = tr)) +
        labs(x = variable, fill = "Treatment") + theme_bw() +
        theme(axis.text.x = element_text(size = 12), axis.text.y = element_text(size = 12),
            axis.title = element_text(size = 16), title = element_text(size = 16),
            legend.title = element_text(size = 12), legend.text = element_text(size = 14),
            strip.text.x = element_text(size = 12))
    return(p)
})

# Move the legend
tmp <- ggplot_gtable(ggplot_build(plot_distrs[[1]]))
leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
legend <- tmp$grobs[[leg]]

plot_distrs <- lapply(plot_distrs, function(x) x + theme(legend.position = "none"))
plot_distrs[[length(plot_distrs) + 1]] <- legend

# Save figure for paper
pdf("../ms/fig/clinical_distr.pdf", width = 8, height = 4)
do.call(grid.arrange, c(plot_distrs, nrow = 2))
dev.off()
```

```
## pdf
##   2
```

```r
# Print figure
do.call(grid.arrange, c(plot_distrs, nrow = 2))
```

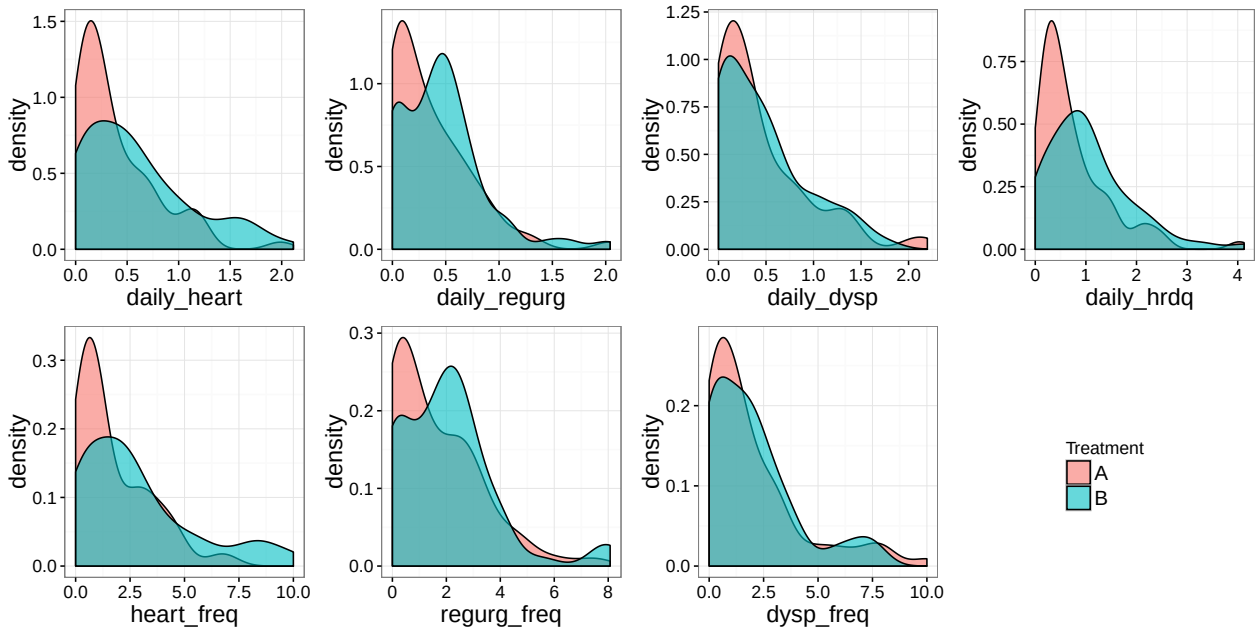

Figure 2: Comparison of the distributions of the seven continuous clinical endpoints during week 2 for the two treatment groups.

```
outcome_summary <- clinical %>% group_by(tr) %>% summarise(mean(daily_heart),
    mean(daily_regurg), mean(daily_dysp), mean(daily_hrdq), mean(heart_freq),
    mean(regurg_freq), mean(dysp_freq)) %>% select(-tr) %>% round(2) %>%
    t()
outcome_sds <- clinical %>% group_by(tr) %>% summarise(sd(daily_heart),
    sd(daily_regurg), sd(daily_dysp), sd(daily_hrdq), sd(heart_freq),
    sd(regurg_freq), sd(dysp_freq)) %>% select(-tr) %>% round(2) %>%
    t()
outcome_summary <- cbind(outcome_summary[, 1], outcome_sds[,
    1], outcome_summary[, 2], outcome_sds[, 2])
colnames(outcome_summary) <- c("Mean A", "SD A", "Mean B", "SD B")
rownames(outcome_summary) <- continuous_vars
print(xtable(outcome_summary, caption = "Mean and standard deviation of each continuous outcome in grou
    include.rownames = TRUE)
```

	Mean A	SD A	Mean B	SD B
daily_heart	0.59	0.46	0.75	0.52
daily_regurg	0.52	0.44	0.58	0.45
daily_dysp	0.56	0.51	0.55	0.49
daily_hrdq	1.11	0.84	1.33	0.86
heart_freq	2.48	2.10	3.51	2.76
regurg_freq	2.30	1.97	2.58	2.16
dysp_freq	2.49	2.48	2.36	2.25

Table 1: Mean and standard deviation of each continuous outcome in groups A and B.

Table 2 shows the correlation between baseline and outcome for each variable. Correlations range between 0.548 and 0.696. This suggests that using the difference in outcome and baseline as the dependent variable in the following models might yield higher power than controlling for baseline as a covariate.

```r
baseline_outcome_corr <- rep(0, length(continuous_vars))
for (i in seq_along(continuous_vars)) {
    col <- continuous_vars[i]
    tmpdata <- reshape_data(col)
    baseline_outcome_corr[i] <- cor(tmpdata$Baseline, tmpdata$Outcome)
}

baseline_outcome_corr <- data.frame(baseline_outcome_corr)
colnames(baseline_outcome_corr) <- "Correlation"
rownames(baseline_outcome_corr) <- continuous_vars


summarytabc <- xtable(baseline_outcome_corr, digits = 3, align = paste0(c("r|",
    rep("p{1in}", 1)), collapse = ""), caption = "Correlation between baseline and outcome, for each co
    label = "tab:clinical_corr")
print(summarytabc, include.rownames = TRUE)
```

	Correlation
daily_heart	0.601
daily_regurg	0.589
daily_dysp	0.572
daily_hrdq	0.562
heart_freq	0.696
regurg_freq	0.548
dysp_freq	0.630

Table 2: Correlation between baseline and outcome, for each continuous endpoint.

Analysis - Primary endpoint

To illustrate the different methods, let's first restrict attention to the primary endpoint, `daily_hrdq`. Throughout this section, we compare two models which both control for site ID. The first model uses the outcome in week two as the dependent variable, controlling for the baseline score in the models with covariates. The second model uses the difference between outcome and baseline score as its dependent variable. The difference in the two models is the way in which we control for the baseline measure.

We compare four testing methods:

1. the usual parametric ANCOVA,
2. a permutation test, where the difference in means is the test statistic and permutations are stratified by site,
3. a variation of this stratified permutation test, using the t-statistic from linear regression instead, and
4. a stratified permutation test in the style of Freedman and Lane (1983), using a linear regression to adjust for covariates.

The permutation tests rely on the `permuter` R package. If you haven't installed it, you should run the following command:

```
devtools::install_github("statlab/permuter")
```

The primary endpoint is `daily_hrdq`, a measure derived from other variables that were measured in the survey. We show summary statistics for the change from baseline for each treatment group below. Based on these statistics, it appears that members of group B had both larger outcomes and larger changes from baseline than group A. We will do two-sided hypothesis tests to assess whether the difference between groups A and B is meaningfully different from zero.

```
daily_hrdq <- reshape_data("daily_hrdq")
summary_stats_hrdq <- cbind(summary(daily_hrdq$Outcome[daily_hrdq$tr ==
    "B"]), summary(daily_hrdq$Outcome[daily_hrdq$tr == "A"]),
    summary(daily_hrdq$difference[daily_hrdq$tr == "B"]), summary(daily_hrdq$difference[daily_hrdq$tr ==
        "A"]))
colnames(summary_stats_hrdq) <- c("B Outcome", "A Outcome", "B Difference",
    "A Difference")
print(xtable(summary_stats_hrdq, caption = "Summary statistics for change in primary
                      endpoint (daily hrdq) for the two treatment groups."),
    include.rownames = TRUE)
```

	B Outcome	A Outcome	B Difference	A Difference
Min.	0.00	0.00	-2.79	-2.81
1st Qu.	0.45	0.25	-0.90	-1.10
Median	0.91	0.47	-0.60	-0.69
Mean	1.04	0.74	-0.57	-0.74
3rd Qu.	1.45	0.95	-0.16	-0.37
Max.	4.13	4.03	1.61	0.91

Table 3: Summary statistics for change in primary endpoint (daily hrdq) for the two treatment groups.

Parametric ANCOVA

```
# method 0: parametric ANCOVA
lm1 <- lm(Outcome ~ Baseline + tr + factor(SITEID), data = daily_hrdq)
daily_hrdq$lm1_resid <- residuals(lm1)
daily_hrdq$lm1_fitted <- fitted(lm1)

lm2 <- lm(difference ~ tr + factor(SITEID), data = daily_hrdq)
daily_hrdq$lm2_resid <- residuals(lm2)
daily_hrdq$lm2_fitted <- fitted(lm2)
```

First, we do a standard parametric ANCOVA using the two models. Under model 1, using the outcome as response and adjusting for the baseline value, treatment has a significant effect with p-value 0.043. Under model 2, the significance goes away (p-value of 0.194). (Note that treatment is still coded as A and B with the character type. The `lm` function is smart enough to know that it should be coded as a factor, where we compare the effect of B relative to A, and coerces the variable type from string to factor behind the scenes. Be very careful when you're doing this! It is generally safer to change types yourself.)

```
print(xtable(summary(aov(lm1))), include.rownames = TRUE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Baseline	1	27.00	27.00	65.16	0.0000
tr	1	1.73	1.73	4.17	0.0432
factor(SITEID)	7	4.50	0.64	1.55	0.1563
Residuals	126	52.20	0.41		

```
print(xtable(summary(aov(lm2))), include.rownames = TRUE)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tr	1	0.92	0.92	1.71	0.1936
factor(SITEID)	7	5.48	0.78	1.46	0.1876
Residuals	127	68.13	0.54		

```
p01 <- daily_hrdq %>% ggplot(aes(x = lm1_fitted, y = lm1_resid)) +
    geom_point() + geom_hline(yintercept = 0, linetype = "dashed") +
    xlab("Fitted Values") + ylab("Residual") + ggtitle("Raw Outcome Model") +
    theme_bw()
p02 <- daily_hrdq %>% ggplot(aes(x = lm2_fitted, y = lm2_resid)) +
    geom_point() + geom_hline(yintercept = 0, linetype = "dashed") +
    xlab("Fitted Values") + ylab("Residual") + ggtitle("Differenced Model") +
    theme_bw()
grid.arrange(p02, p01, nrow = 1)
```

Figure 3 shows the fitted values versus residuals for the two ANCOVA models. The residuals of the raw outcome model appear to increase in magnitude with the fitted value, casting doubt on the assumption that the error variance is constant. The plot reveals that the differenced model only yields a few predicted values, but does not seem to indicate heteroskedasticity.

Stratified permutation test

Suppose each individual, $i = 1, \ldots, N$ has two potential outcomes $(Y_i(A), Y_i(B))$ that indicate their response to treatments A and B, respectively. We randomly assign treatment to individuals; treatment determines which of $Y_i(A)$ and $Y_i(B)$ we observe. We are unable to observe both. Suppose we wish to test the null

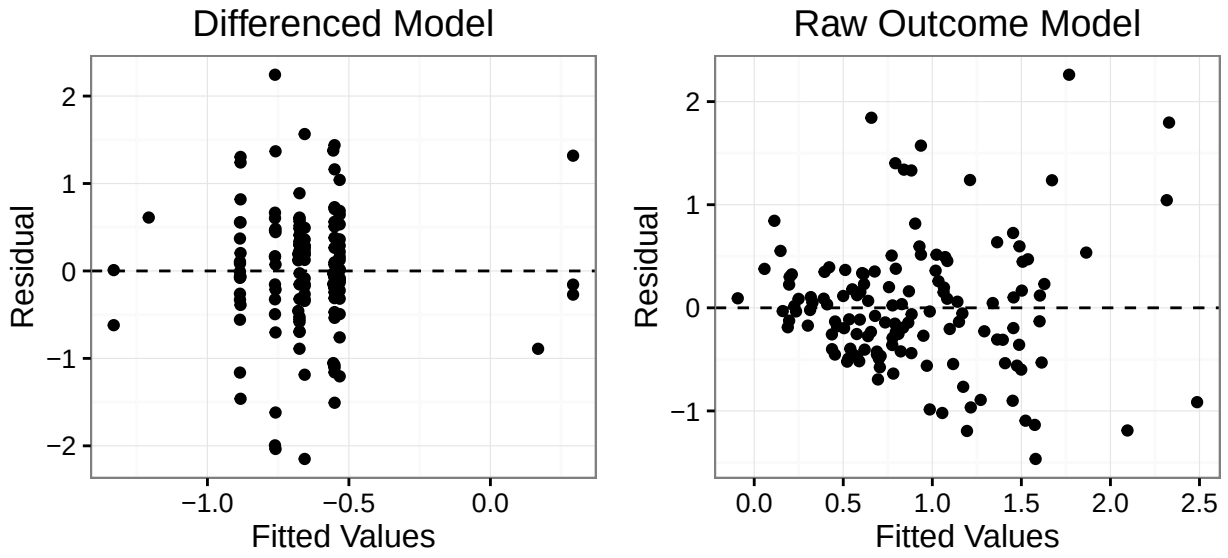

Figure 3: Residual plots for the ANCOVA model of daily hrdq.

hypothesis that individual by individual, treatment has no effect. This is referred to as the "sharp null" hypothesis:

$$H_0 : Y_i(A) = Y_i(B), i = 1, \ldots, N.$$

Then, whether individual i received A or B amounts to an arbitrary label. Once we observe their response under one treatment, we know what it would have been under the other; namely, it would be the same.

Treatment was completely randomized at each site, independently across sites. We assume that drop-out from the study was independent of treatment assignment (which is true if treatment was blinded and assigned at random). This will be a conditional hypothesis test: we condition on the number of individuals who received A and B at each site. This conditioning asserts that any assignment of treatments that preserves the number of treated and controls at each site is valid and has an equal probability of occurring. Therefore, using this principle of equal probabilities and by imputing the unobserved potential outcomes assuming that the null hypothesis is true, we can compute the permutation distribution of any statistic under the null hypothesis.

We compare two test statistics: the difference in mean outcomes for group B and mean outcomes for group A, and this difference of means within each individual site, aggregated over sites by taking the sum of their absolute values. The first statistic is more comparable to what one would obtain from an ANCOVA and it is readily interpretable. It does not directly account for the stratification by site, so variation between sites may be hidden. The second statistic is useful for testing the two-sided alternative hypothesis that treatment has *some* nonzero effect. It may be more powerful than the simple difference in means if the effect of the drug varies across sites. For instance, if the drug has a positive effect at one site but a negative effect at another site, this test statistic would be large. If we used the simple difference in means without accounting for sites, then the positive effect and negative effect may cancel each other out; the difference in means will be attenuated toward zero.

```
# method 1 : do permutation of differences
observed_diff_means1 <- mean(daily_hrdq[daily_hrdq$tr == "B",
    ]$Outcome) - mean(daily_hrdq[daily_hrdq$tr == "A", ]$Outcome)
diff_means_distr1 <- stratified_two_sample(group = daily_hrdq$tr,
    response = daily_hrdq$Outcome, stratum = daily_hrdq$SITEID,
    reps = 10000)
diff_means_pvalue1 <- t2p(observed_diff_means1, diff_means_distr1,
```

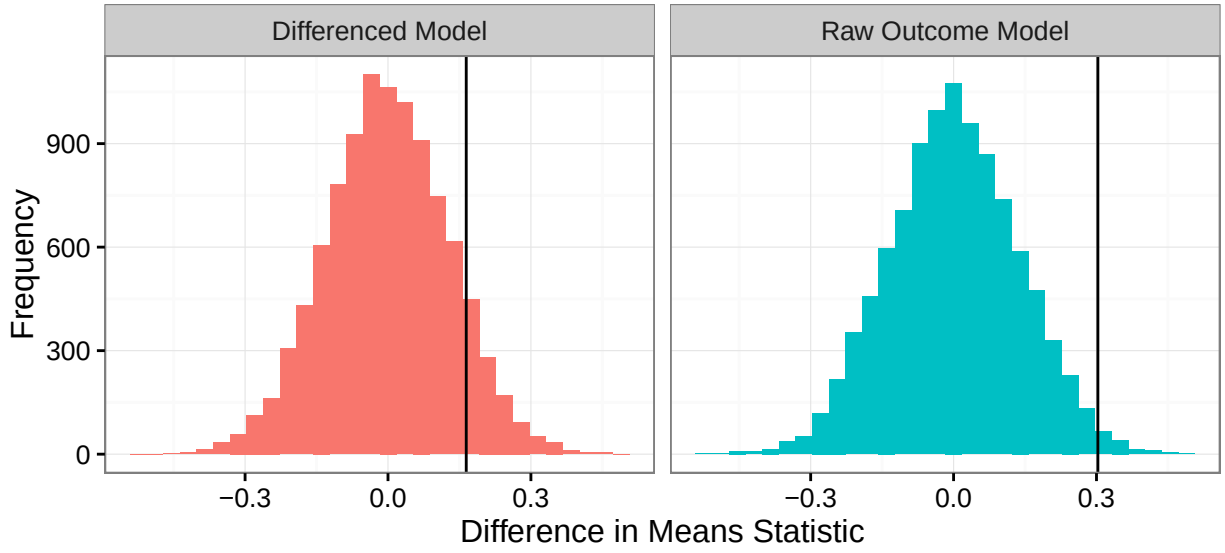

Figure 4: Permutation distribution of the difference of means for daily hrdq

```
    alternative = "two-sided")


observed_diff_means2 <- mean(daily_hrdq[daily_hrdq$tr == "B",
    ]$difference) - mean(daily_hrdq[daily_hrdq$tr == "A", ]$difference)
diff_means_distr2 <- stratified_two_sample(group = daily_hrdq$tr,
    response = daily_hrdq$difference, stratum = daily_hrdq$SITEID,
    reps = 10000)
diff_means_pvalue2 <- t2p(observed_diff_means2, diff_means_distr2,
    alternative = "two-sided")


data.frame(perm = c(diff_means_distr1, diff_means_distr2), model = c(rep("Raw Outcome Model",
    length(diff_means_distr1)), rep("Differenced Model", length(diff_means_distr2))),
    xintercept = c(rep(observed_diff_means1, length(diff_means_distr1)),
        rep(observed_diff_means2, length(diff_means_distr2)))) %>%
    ggplot(aes(x = perm, fill = model)) + geom_histogram() +
    facet_grid(~model) + geom_vline(aes(xintercept = xintercept)) +
    theme_bw() + labs(x = "Difference in Means Statistic", y = "Frequency") +
    theme(legend.position = "none")
```

```
obs_diff_means_bystrata1 <- sum(abs(within_group_mean(group = daily_hrdq$tr,
    response = daily_hrdq$Outcome, stratum = daily_hrdq$SITEID,
    groups = unique(daily_hrdq$tr), strata = unique(daily_hrdq$SITEID))))
diff_means_distr_bystrata1 <- stratified_two_sample(group = daily_hrdq$tr,
    response = daily_hrdq$Outcome, stratum = daily_hrdq$SITEID,
    stat = "mean_within_strata", reps = 10000)
diff_means_bystrata_pvalue1 <- t2p(obs_diff_means_bystrata1,
    diff_means_distr_bystrata1, alternative = "two-sided")


obs_diff_means_bystrata2 <- sum(abs(within_group_mean(group = daily_hrdq$tr,
    response = daily_hrdq$difference, stratum = daily_hrdq$SITEID,
```

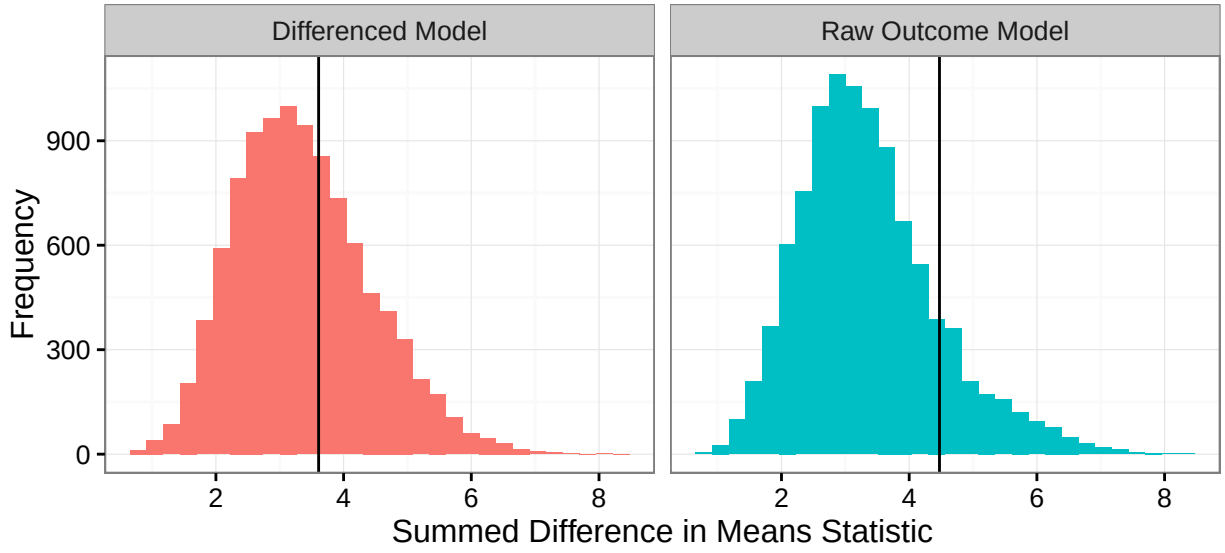

Figure 5: Permutation distribution of the difference of means within strata, summed across strata, for daily hrdq

```
    groups = unique(daily_hrdq$tr), strata = unique(daily_hrdq$SITEID))))
diff_means_distr_bystrata2 <- stratified_two_sample(group = daily_hrdq$tr,
    response = daily_hrdq$difference, stratum = daily_hrdq$SITEID,
    stat = "mean_within_strata", reps = 10000)
diff_means_bystrata_pvalue2 <- t2p(obs_diff_means_bystrata2,
    diff_means_distr_bystrata2, alternative = "two-sided")


data.frame(perm = c(diff_means_distr_bystrata1, diff_means_distr_bystrata2),
    model = c(rep("Raw Outcome Model", length(diff_means_distr_bystrata1)),
        rep("Differenced Model", length(diff_means_distr_bystrata2))),
    xintercept = c(rep(obs_diff_means_bystrata1, length(diff_means_distr_bystrata1)),
        rep(obs_diff_means_bystrata2, length(diff_means_distr_bystrata2)))) %>%
    ggplot(aes(x = perm, fill = model)) + geom_histogram() +
    facet_grid(~model) + geom_vline(aes(xintercept = xintercept)) +
    theme_bw() + labs(x = "Summed Difference in Means Statistic",
    y = "Frequency") + theme(legend.position = "none")
```

The empirical permutation distributions are computed by randomly permuting treatment assignments within each site 10^4 times and computing the statistic for each permutation. A vertical line indicates the observed value of the test statistic. The p-value is two times the area of the permutation distribution to the right of the observed value; it is the probability of observing a value at least as extreme (falling that far away from the center of the distribution) as the observed value. Figure 4 shows the permutation distributions of the simple difference in means for the two models. The p-values for the differenced model and raw outcome model are 0.2 and 0.023, respectively. Figure 5 shows the permutation distribution of the absolute value of the difference of means within strata, summed across strata. The p-values for the differenced model and raw outcome model are 0.75 and 0.288, respectively.

There are several points to notice:

- The second test is far less powerful than the first. This is likely because some sites had very few patients. These sites have correspondingly few unique permutations, and so they often contribute the same amount to the summed difference in means statistic. It is impossible to detect a difference at small

significance levels when there are only 3 patients.
- The summed difference of means only makes sense for testing against a two-sided alternative, since taking the absolute value removes the sign of the effect. The permutation distribution is strictly non-negative and skewed to the right.
- This statistic is less interpretable than the difference in means because it doesn't correspond directly to any feature of the distribution of outcomes.

Since the summed difference in means statistic has such low power, we abandon it in the rest of the analyses.

Covariate-adjusted permutation test

We would like to test for a difference in outcomes between the treatment groups, but control for other covariates. In particular, when the outcome is the average response during the second week of follow-up, we would like to control for the average response during the first week of follow-up and the location. When the outcome is the difference in responses between the second and first weeks, we would just like to control for location. We will use the approximate permutation test derived by Freedman and Lane (1983) to do so.

Let Y denote the response, Z denote the treatment indicator (1 if the person gets treatment B, 0 if they get A), and X denote a covariate which may be correlated with Y. We may write the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Under the sharp null H_0, Z has no effect on Y. In other words, the null hypothesis is that $\beta_2 = 0$. In a randomized experiment like this, several variables are exchangeable. We will show two different permutation tests that exploit exchangeability of different variables.

Standard linear model

Treatment was assigned at random within each site. This ensures that Z and ε are statistically independent, conditional on site. Therefore, we may conduct a test by permuting treatment assignments Z within site, independently across sites, and calculating a test statistic for each such permutation. We choose to use the t statistic from the linear model as the test statistic.

```
# method 2 : linear model

observed_t1 <- summary(lm1)[["coefficients"]]["trB", "t value"]
lm1_t_distr <- replicate(10000, {
    daily_hrdq$tr_perm <- permute_within_groups(daily_hrdq$tr,
        daily_hrdq$SITEID)
    lm1_perm <- lm(Outcome ~ Baseline + tr_perm + factor(SITEID),
        data = daily_hrdq)
    summary(lm1_perm)[["coefficients"]]["tr_permB", "t value"]
})

lm_pvalue_1 <- t2p(observed_t1, lm1_t_distr, alternative = "two-sided")


observed_t2 <- summary(lm2)[["coefficients"]]["trB", "t value"]
lm2_t_distr <- replicate(10000, {
    daily_hrdq$tr_perm <- permute_within_groups(daily_hrdq$tr,
        daily_hrdq$SITEID)
    lm2_perm <- lm(difference ~ tr_perm + factor(SITEID), data = daily_hrdq)
    summary(lm2_perm)[["coefficients"]]["tr_permB", "t value"]
```

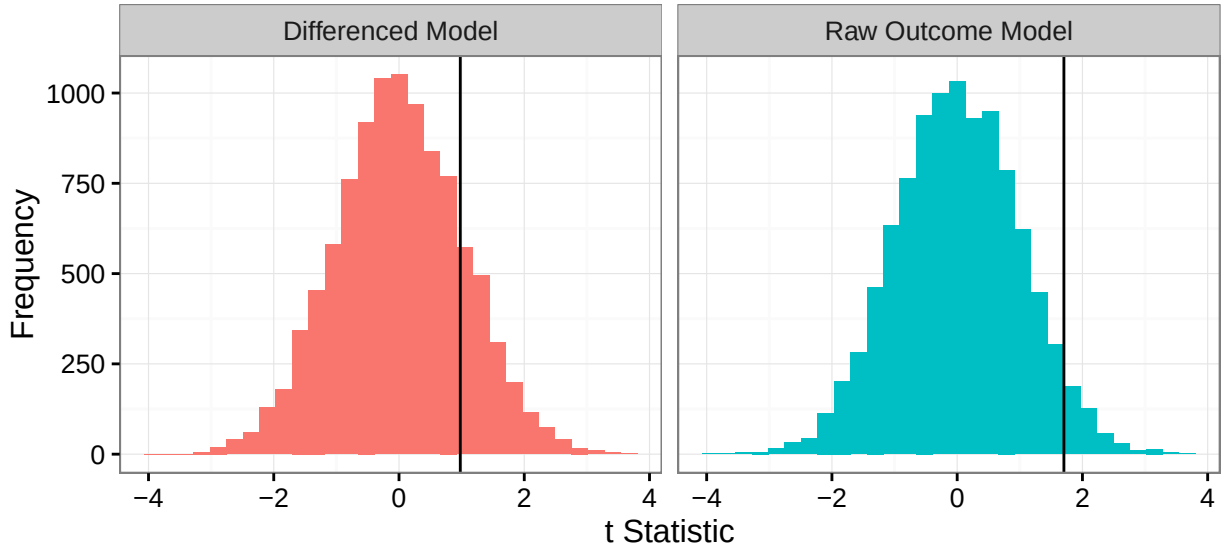

Figure 6: Permutation distribution of the stratified permutation test between daily hrdq and treatment after controlling for covariates in a linear model.

```
})

lm_pvalue_2 <- t2p(observed_t2, lm2_t_distr, alternative = "two-sided")


data.frame(perm = c(lm1_t_distr, lm2_t_distr), model = c(rep("Raw Outcome Model",
    length(lm1_t_distr)), rep("Differenced Model", length(lm2_t_distr))),
    xintercept = c(rep(observed_t1, length(lm1_t_distr)), rep(observed_t2,
        length(lm2_t_distr)))) %>% ggplot(aes(x = perm, fill = model)) +
    geom_histogram() + facet_grid(~model) + geom_vline(aes(xintercept = xintercept)) +
    theme_bw() + labs(x = "t Statistic", y = "Frequency") + theme(legend.position = "none")
```

The p-values for the differenced model and raw outcome model are 0.335 and 0.087, respectively.

Freedman-Lane test of residuals

Let's take an alternative view of the problem. We still write $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$. However, now, we do not treat the ε as random. They are simply defined to be the difference between Y and the data's linear projection onto the plane $\beta_0 + \beta_1 X + \beta_2 Z$.

If the null hypothesis is true, then $\varepsilon = Y - \beta_0 - \beta_1 X$. Therefore, we may estimate the errors $\hat{\varepsilon}$ by $Y - \hat{Y}$, where $\hat{Y}$ is obtained by regressing Y on X but not Z. Assuming that the $\hat{\varepsilon}$ are "close" to ε, then these estimated $\hat{\varepsilon}$ are approximately exchangeable within sites.

We construct a permutation distribution using several steps:

1. Estimate $\hat{\varepsilon}$ by $Y - \hat{\beta}_0 - \hat{\beta}_1 X$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by regressing Y on X.
2. Construct permuted errors $\hat{\varepsilon}^\pi$ by permuting the $\hat{\varepsilon}$ within sites.
3. Construct permuted responses $Y^\pi = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$.
4. Regress Y^π on X and Z. The test statistic is the t-statistic for the coefficient of Z.

For this dataset, we are guaranteed that treatment Z is independent of ε, as it is randomized within site. However, the randomization does not guarantee that ε is independent baseline measure, conditional on site.

We check this associations using residual plots in Figure 7. The distribution of residuals varies a bit across sites, but they all appear roughly centered around 0.

```
obs_corr <- daily_hrdq %>% group_by(SITEID) %>% summarise(n = n(),
    Correlation = cor(lm1_resid, Baseline))
corr_pvalues <- rep(0, 8)

for (i in unique(daily_hrdq$SITEID)) {
    tmp <- daily_hrdq %>% filter(SITEID == i)
    corr_distr <- replicate(10000, {
        cor(permute(tmp$lm1_resid), tmp$Baseline)
    })
    corr_pvalues[i] <- t2p(unlist(obs_corr[i, "Correlation"]),
        corr_distr, alternative = "two-sided")
}

fisher_combined_corr <- pchisq(fisher(corr_pvalues), df = 2 *
    length(corr_pvalues), lower.tail = FALSE)

obs_corr <- obs_corr %>% mutate(`P-value` = corr_pvalues)
print(xtable(obs_corr, digits = 2, caption = "Permutation test $p$-values for the correlation between ba
```

SITEID	n	Correlation	P-value
1	38	0.32	0.06
2	7	-0.18	0.73
3	16	-0.49	0.06
4	9	-0.34	0.39
5	4	0.92	0.41
6	21	-0.03	0.94
7	38	-0.13	0.43
8	3	-0.92	0.35

Table 4: Permutation test p-values for the correlation between baseline daily hrdq and residuals.

By construction of the linear model, the correlation between the baseline measure and the residuals is zero. However, we would like to check formally that ε is independent baseline measure for each site, using the residuals as a proxy for ε. To do so, we conduct independent permutation tests within each stratum using the Pearson correlation as the statistic. Table 5 shows the results of this test: while the correlation between baseline and residuals appears high, it is not statistically significant at the 5% level in any stratum. The Fisher's combined p-value for the 8 strata is 0.24, so there is no evidence that the independence assumption is violated.

```
# method 3 : freedman lane perm residuals

lm1_no_tr <- lm(Outcome ~ Baseline + factor(SITEID), data = daily_hrdq)
lm1_resid <- residuals(lm1_no_tr)
lm1_yhat <- fitted(lm1_no_tr)
observed_t1 <- summary(lm1)[["coefficients"]]["trB", "t value"]
lm1_t_distr <- replicate(10000, {
    lm1_resid_perm <- permute_within_groups(lm1_resid, daily_hrdq$SITEID)
    daily_hrdq$response_fl <- lm1_yhat + lm1_resid_perm
    lm1_perm <- lm(response_fl ~ Baseline + tr + factor(SITEID),
        data = daily_hrdq)
    summary(lm1_perm)[["coefficients"]]["trB", "t value"]
})
```

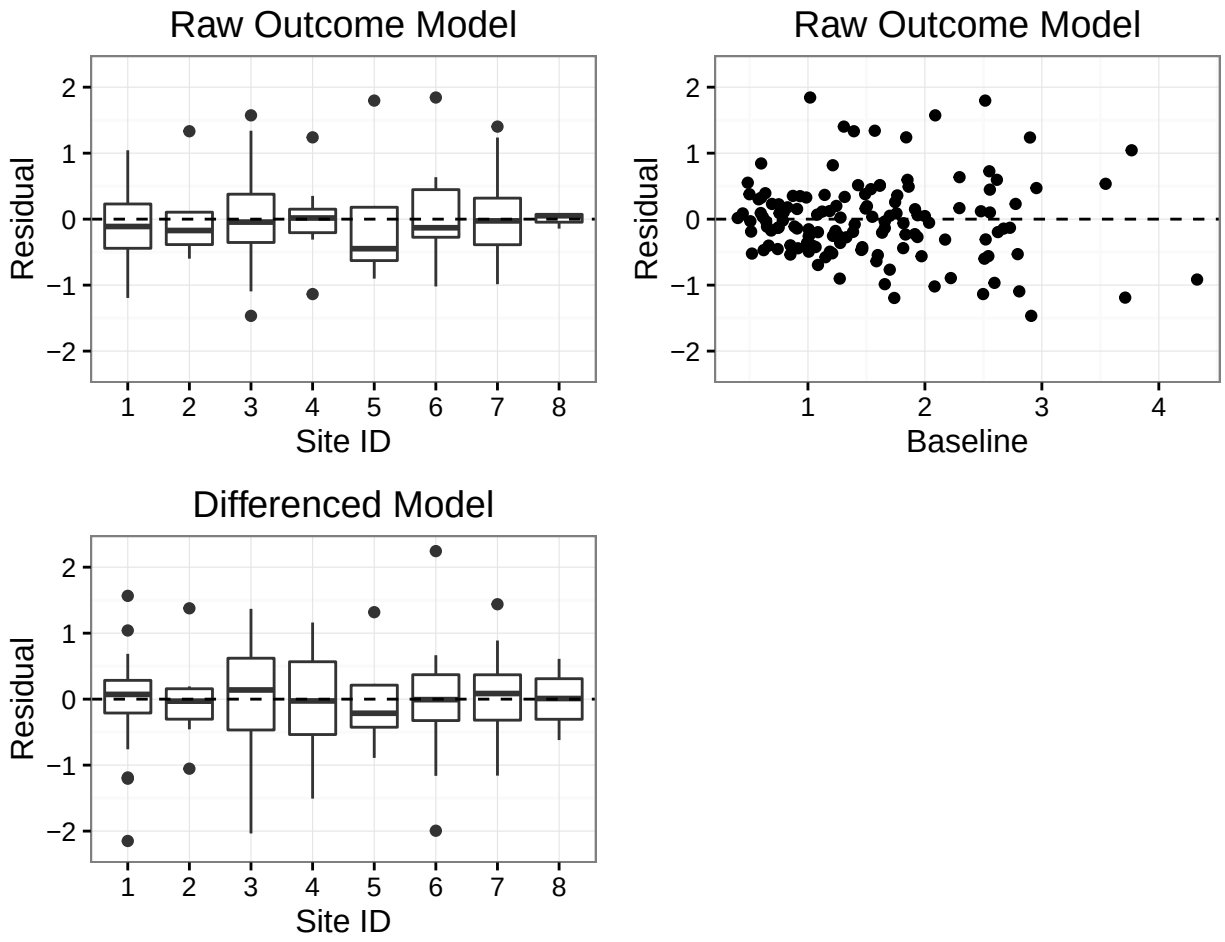

Figure 7: Residual plots of the linear regression of daily hrdq on treatment and covariates. These are done to check that treatment and covariates are uncorrelated with the regression errors for the Freedman and Lane style permutation test.

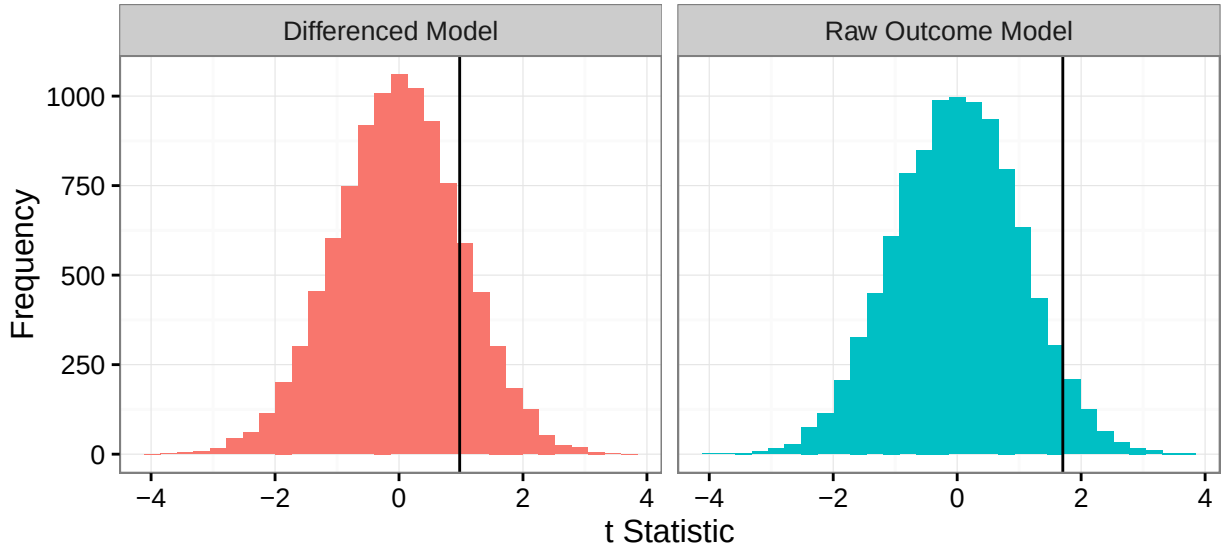

Figure 8: Permutation distribution of the Freedman-Lane tests of correlation between daily hrdq and treatment after controlling for covariates.

```
fl_pvalue_1 <- t2p(observed_t1, lm1_t_distr, alternative = "two-sided")



lm2_no_tr <- lm(difference ~ factor(SITEID), data = daily_hrdq)
lm2_resid <- residuals(lm2_no_tr)
lm2_yhat <- fitted(lm2_no_tr)
observed_t2 <- summary(lm2)[["coefficients"]]["trB", "t value"]
lm2_t_distr <- replicate(10000, {
    lm2_resid_perm <- permute_within_groups(lm2_resid, daily_hrdq$SITEID)
    daily_hrdq$response_fl <- lm2_yhat + lm2_resid_perm
    lm2_perm <- lm(response_fl ~ tr + factor(SITEID), data = daily_hrdq)
    summary(lm2_perm)[["coefficients"]]["trB", "t value"]
})

fl_pvalue_2 <- t2p(observed_t2, lm2_t_distr, alternative = "two-sided")


data.frame(perm = c(lm1_t_distr, lm2_t_distr), model = c(rep("Raw Outcome Model",
    length(lm1_t_distr)), rep("Differenced Model", length(lm2_t_distr))),
    xintercept = c(rep(observed_t1, length(lm1_t_distr)), rep(observed_t2,
        length(lm2_t_distr)))) %>% ggplot(aes(x = perm, fill = model)) +
    geom_histogram() + facet_grid(~model) + geom_vline(aes(xintercept = xintercept)) +
    theme_bw() + labs(x = "t Statistic", y = "Frequency") + theme(legend.position = "none")
```

The p-values for the differenced model and raw outcome model are 0.327 and 0.097, respectively.

Summary

Table 4 shows the p-values for each of the five methods and the two methods of controlling for the baseline. For the models using the raw outcome measure as the dependent variable, the effect of `daily_hrdq` is significant at the 0.05 level in two of the five tests and is significant at the 0.1 level in four of the five tests. It is never significant in the models using the difference from baseline to outcome. As expected, the stratified permutation test using the sum of differences in means across sites has low power.

	Differences	Outcome
Parametric ANCOVA	0.194	0.043
Unadjusted permutation	0.200	0.023
Unadjusted permutation (summed across strata)	0.750	0.288
Linear regression permutation	0.335	0.087
Residual permutation	0.327	0.097

Table 5: Comparison of p-values for two measures (average outcome during treatment vs. difference of average outcome and average baseline) of the primary endpoint.

Analysis - Secondary endpoints

We restrict our attention to the model using treatment measurement and controlling for the baseline, as this is what RB's original analysis does. We run this procedure for all seven continuous endpoints. Table 6 shows the results.

Figure 9 displays the residuals versus fitted values for the linear model using each response variable. The figures cast doubt on the assumption of constant error variance in the linear model: they indicate that the variance of the residuals increases with the fitted values and the distribution of residuals may have some right skew. The assumption of normally distributed errors with constant variance is important for the parametric analysis of the ANCOVA model, but does not apply to the permutation tests. Across all tests, the parametric ANCOVA p-value is smaller than the linear model permutation test p-values. Perhaps this is partly due to the heteroskedasticity: the permutation tests make no distributional assumptions, so may be less sensitive to large outcomes or residuals.

```
set.seed(919547773)  # Generated on Random.org Timestamp: 2016-11-09 15:39:29 UTC

continuous_vars <- c("heart_freq", "regurg_freq", "dysp_freq",
    "daily_heart", "daily_regurg", "daily_hrdq", "daily_dysp")

tests <- c("ANCOVA", "Stratified perm", "LM perm", "Residual perm")
pvalues_table_contin <- as.data.frame(matrix(NA, nrow = length(continuous_vars),
    ncol = 4))
baseline_outcome_corr <- rep(0, length(continuous_vars))

for (i in seq_along(continuous_vars)) {
    col <- continuous_vars[i]
    tmpdata <- reshape_data(col)
    baseline_outcome_corr[i] <- cor(tmpdata$Baseline, tmpdata$Outcome)

    lm1 <- lm(Outcome ~ Baseline + tr + factor(SITEID), data = tmpdata)
    pvalues_table_contin[i, 1] <- summary(aov(lm1))[[1]]["tr",
        "Pr(>F)"]

    observed_diff_means <- mean(tmpdata[tmpdata$tr == "B", ]$Outcome) -
        mean(tmpdata[tmpdata$tr == "A", ]$Outcome)
    diff_means_distr <- stratified_two_sample(group = tmpdata$tr,
        response = tmpdata$Outcome, stratum = tmpdata$SITEID,
        reps = 10000)
    pvalues_table_contin[i, 2] <- t2p(observed_diff_means, diff_means_distr,
        alternative = "two-sided")

    observed_t <- summary(lm1)[["coefficients"]]["trB", "t value"]
    lm1_t_distr <- replicate(10000, {
        tmpdata$tr_perm <- permute_within_groups(tmpdata$tr,
            tmpdata$SITEID)
        lm1_perm <- lm(Outcome ~ Baseline + tr_perm + factor(SITEID),
            data = tmpdata)
        summary(lm1_perm)[["coefficients"]]["tr_permB", "t value"]
    })
    pvalues_table_contin[i, 3] <- t2p(observed_t, lm1_t_distr,
        alternative = "two-sided")

    lm_no_tr <- lm(Outcome ~ Baseline + factor(SITEID), data = tmpdata)
```

```
    lm_resid <- residuals(lm_no_tr)
    lm_yhat <- fitted(lm_no_tr)
    observed_t <- summary(lm1)[["coefficients"]]["trB", "t value"]
    lm_t_distr <- replicate(10000, {
        lm_resid_perm <- permute_within_groups(lm_resid, tmpdata$SITEID)
        tmpdata$response_fl <- lm_yhat + lm_resid_perm
        lm_perm <- lm(response_fl ~ Baseline + tr + factor(SITEID),
            data = tmpdata)
        summary(lm_perm)[["coefficients"]]["trB", "t value"]
    })

    pvalues_table_contin[i, 4] <- t2p(observed_t, lm_t_distr,
        alternative = "two-sided")
}
```

```
pplot <- vector("list", 7)
for (i in seq_along(continuous_vars)) {
    col <- continuous_vars[i]
    tmpdata <- reshape_data(col)
    lm1 <- lm(Outcome ~ Baseline + tr + factor(SITEID), data = tmpdata)
    tmpdata <- tmpdata %>% mutate(Residuals = residuals(lm1),
        Fitted = fitted(lm1))
    pplot[[i]] <- ggplot(tmpdata, aes(x = Fitted, y = Residuals)) +
        geom_point() + ggtitle(continuous_vars[i]) + theme_bw()
}
do.call(grid.arrange, c(pplot, nrow = 2))
```

	ANCOVA	Stratified perm	LM perm	Residual perm
heart_freq	0.035	0.003	0.070	0.074
regurg_freq	0.136	0.157	0.231	0.221
dysp_freq	0.565	0.925	0.592	0.579
daily_heart	0.032	0.010	0.062	0.069
daily_regurg	0.142	0.195	0.250	0.243
daily_hrdq	0.043	0.033	0.086	0.088
daily_dysp	0.582	0.803	0.686	0.699

Table 6: Comparison of p-values from four tests, for each continuous endpoint.

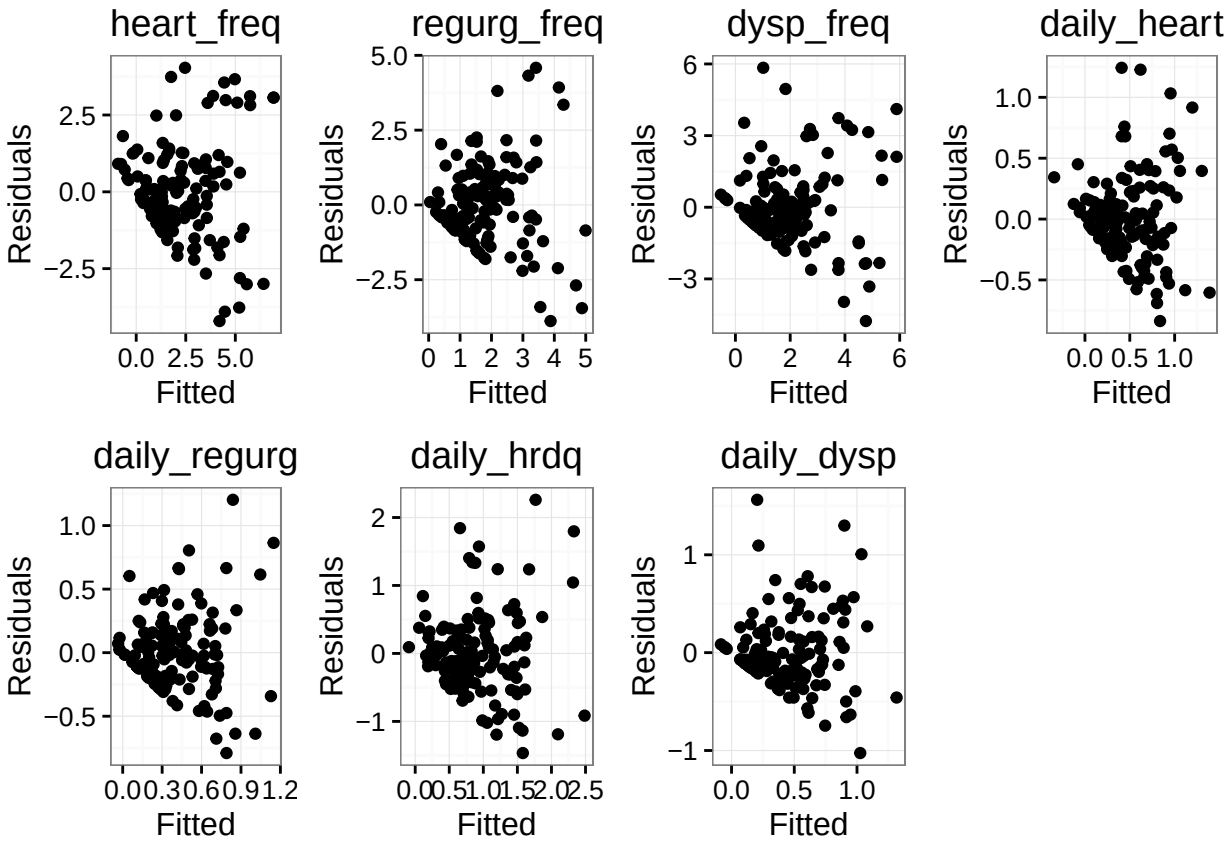

Figure 9: Residual plots for the linear regression of outcome on baseline, treatment, and stratum ID for each response variable.