Clinical Trial Data Analysis

Kellie Ottoboni

2017-02-01

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.

```
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)
```

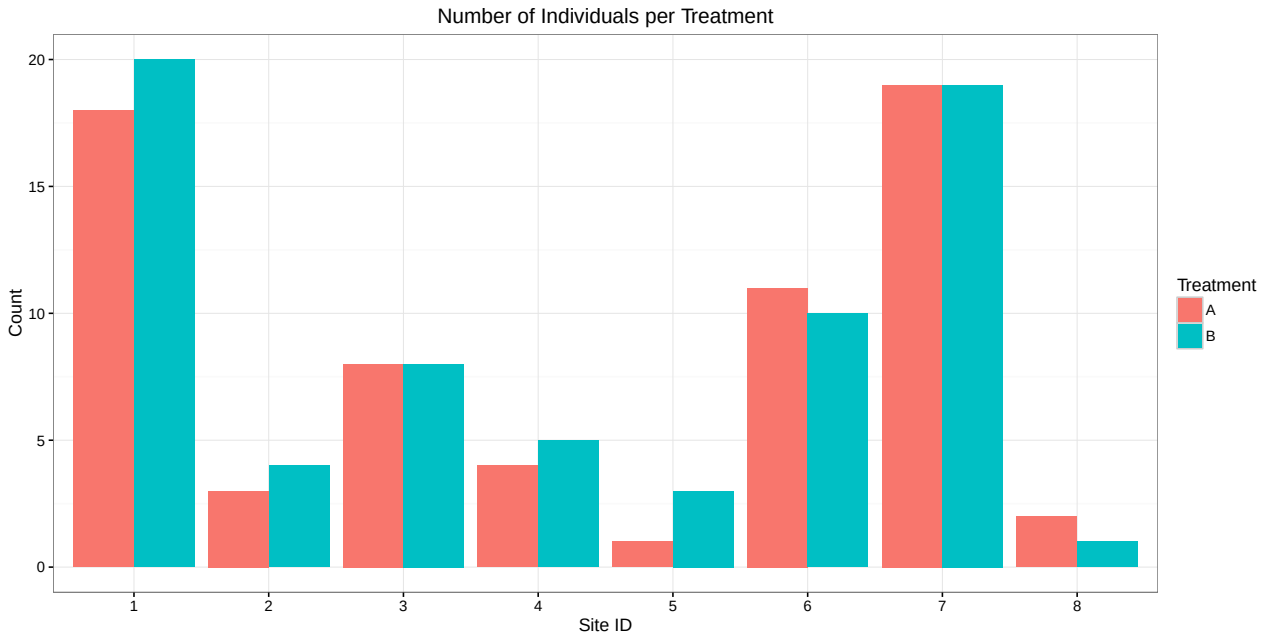

Figure 1: Number of subjects in each treatment group, by site ID.

```
    return(cleaned)
}
```

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

```
# 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)
```

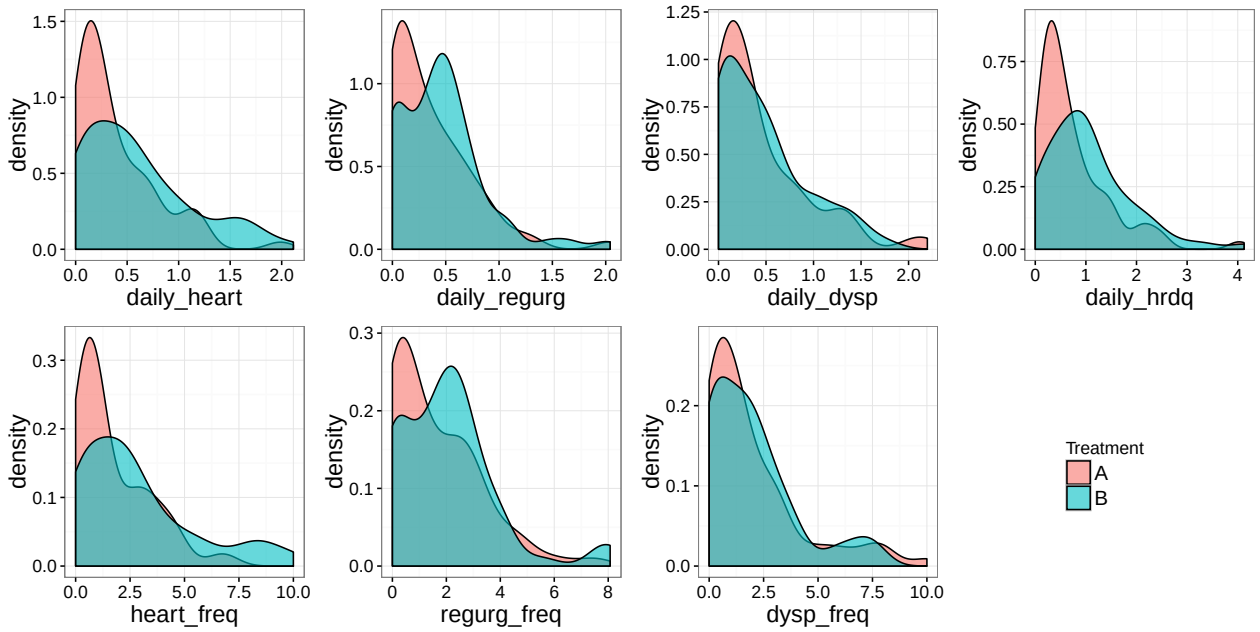

Figure 2: Comparison of the distributions of the seven continuous clinical endpoints during week 2 for the two treatment groups.

```
do.call(grid.arrange, c(plot_distrs, nrow = 2))
dev.off()
```

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

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

The primary endpoint is `daily_hrdq`, a measure derived from other variables that were measured in the survey. We show summary statistics for this outcome for each treatment group below.

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

	B	A
Min.	-2.79	-2.81
1st Qu.	-0.90	-1.10
Median	-0.60	-0.69
Mean	-0.57	-0.74
3rd Qu.	-0.16	-0.37
Max.	1.61	0.91

Table 1: Summary statistics for change in primary endpoint (daily hrdq) for the two treatment groups.

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 how 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.

Parametric ANCOVA

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. Under model 2, the significance goes away. (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.)

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

```
daily_hrdq$lm1_resid <- residuals(lm1)

lm2 <- lm(difference ~ tr + factor(SITEID), data = daily_hrdq)
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		

```
daily_hrdq$lm2_resid <- residuals(lm2)
```

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

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.

Figure 3 shows the permutation distributions of the simple difference in means for the two models. Figure 4 shows the permutation distribution of the absolute value of the difference of means within strata, summed across strata. These empirical permutation distributions are computed by randomly permuting treatment assignments within each site 10^4 times and computing the statistic for each permutation. The vertical line shows 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. 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.

```
# 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,
    alternative = "two-sided")
```

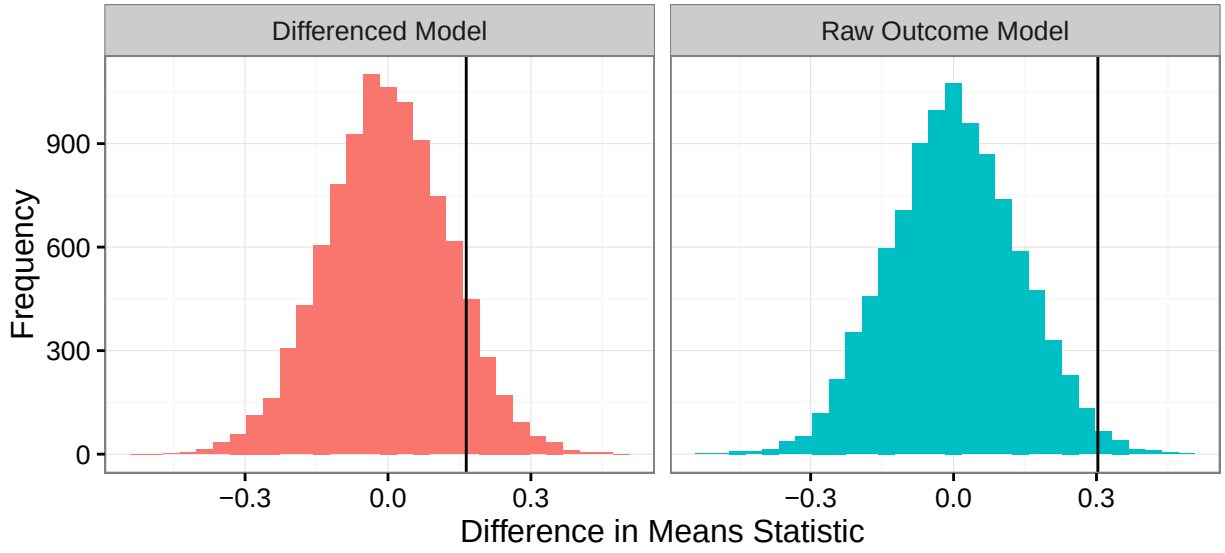

Figure 3: Permutation distribution of the difference of means for daily hrdq

```
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$SITEID,
    groups = unique(daily_hrdq$tr), strata = unique(daily_hrdq$SITEID))))
```

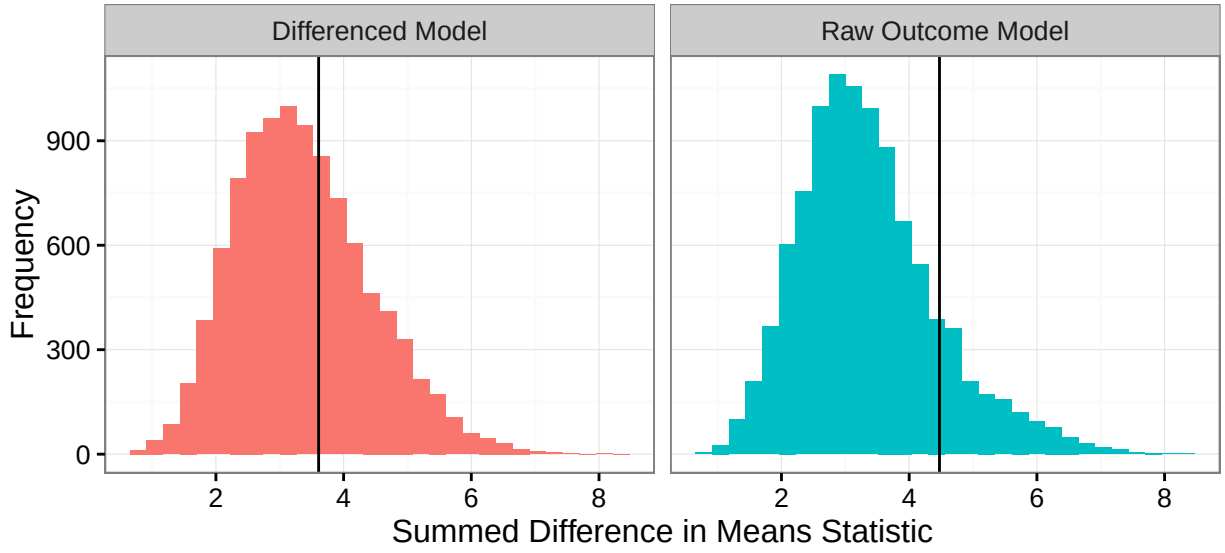

Figure 4: Permutation distribution of the difference of means within strata, summed across strata, for daily hrdq

```
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")
```

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"]
})

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")
```

Linear model 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. The ε are orthogonal to X and Z. Assuming that the $\hat{\varepsilon}$ are "close" to ε, thenthese estimated $\hat{\varepsilon}$ are approximately exchangeable within sites.

We construct a permutation distribution using several steps:

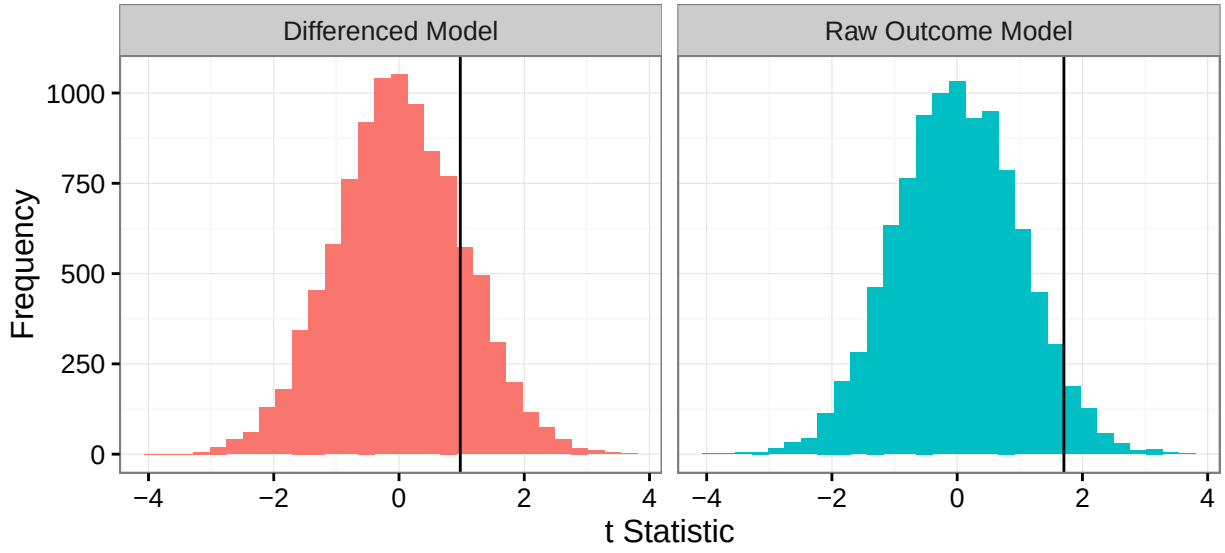

Figure 5: Permutation distribution of the stratified permutation test between daily hrdq and treatment after controlling for covariates in a linear model.

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. We check these associations using residual plots in Figure 6. Indeed, the distribution of residuals looks nearly equal between treatment groups in both models. The distribution of residuals varies a bit across sites, but they all appear roughly centered around 0. Sites 4 and 5 have larger variance in the differenced model. This should not be an issue since we permute treatments within sites, independently across sites. (In general, when treatment is not randomly assigned, this condition is necessary for the permutation test to be valid.)

```
# 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"]
})

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)
```

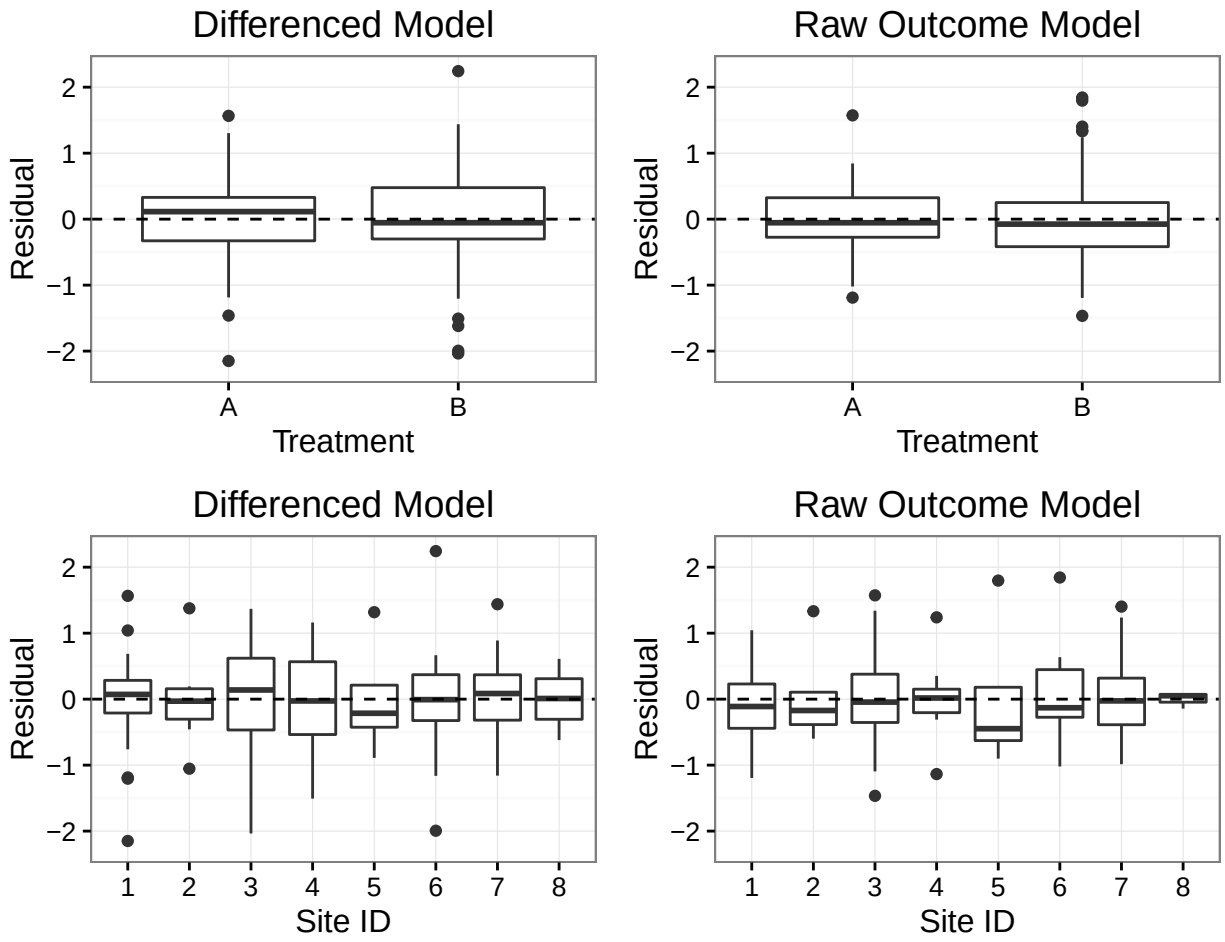

Figure 6: Residual plots of the linear regression of daily hrdq on treatment and covariates. These are done to check that treatment and covariates are uncorrelated with the regression errors for the Freedman and Lane style permutation test.

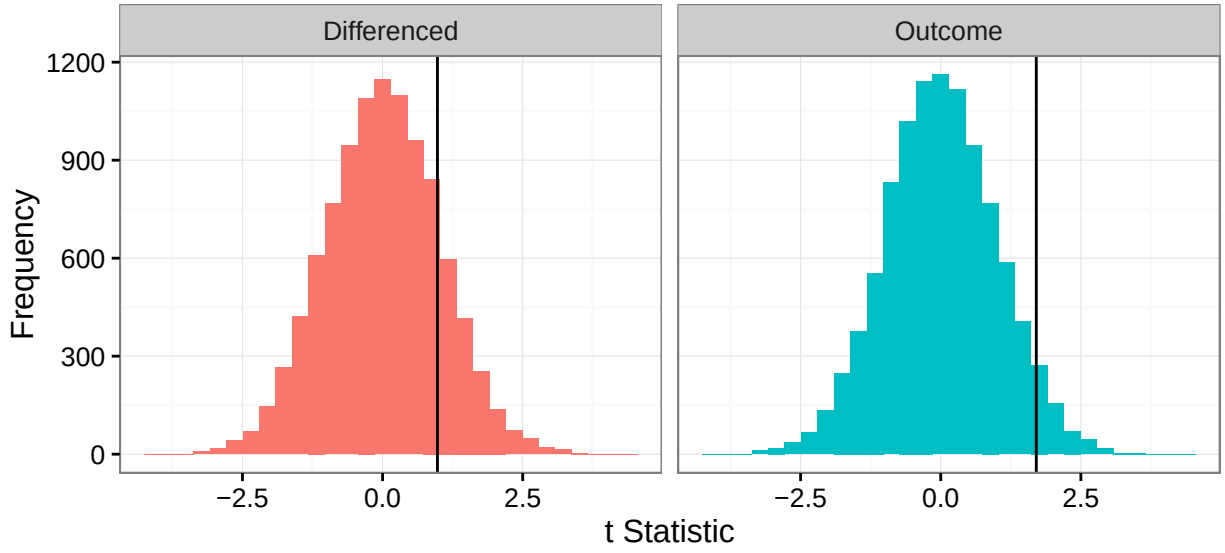

Figure 7: Permutation distribution of the Freedman-Lane tests of correlation between daily hrdq and treatment after controlling for covariates.

```
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("Outcome",
    length(lm1_t_distr)), rep("Differenced", 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")
```

Table 2 shows the p-values for each of the four methods and the two models. 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.

```
outcome_pvalues <- c(summary(aov(lm1))[[1]]["trB", "Pr(>F)"],
    diff_means_pvalue1, diff_means_bystrata_pvalue1, lm_pvalue_1,
    fl_pvalue_1)
differenced_pvalues <- c(summary(aov(lm2))[[1]]["trB", "Pr(>F)"],
    diff_means_pvalue2, diff_means_bystrata_pvalue2, lm_pvalue_2,
    fl_pvalue_2)
pvalues_table <- cbind(differenced_pvalues, outcome_pvalues)
```

```
tests <- c("Parametric ANCOVA", "Unadjusted permutation", "Unadjusted permutation (summed across strata)",
    "Linear regression permutation", "Residual permutation")
rownames(pvalues_table) <- tests
colnames(pvalues_table) <- c("Differences", "Outcome")

print(xtable(pvalues_table, digits = 3, caption = "Comparison of $p$-values for two
             measures (average outcome during treatment
             vs. difference of average outcome and average
             baseline) of the primary endpoint."),
    include.rownames = TRUE)
```

	Differences	Outcome
Parametric ANCOVA		
Unadjusted permutation	0.200	0.023
Unadjusted permutation (summed across strata)	0.750	0.288
Linear regression permutation	0.335	0.087
Residual permutation	0.341	0.094

Table 2: 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 3 shows the results.

```
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))
i <- 0

for (col in continuous_vars) {
    i <- i + 1
    tmpdata <- reshape_data(col)
    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 = 1000)
    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(1000, {
        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(1000, {
        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")
}
```

```
rownames(pvalues_table_contin) <- continuous_vars
colnames(pvalues_table_contin) <- tests


summarytab <- xtable(pvalues_table_contin, digits = 3, align = paste0(c("r|",
    rep("p{1in}", ncol(pvalues_table_contin))), collapse = ""),
    caption = "Comparison of $p$-values from four tests, for each continuous endpoint.",
    label = "tab:clinical_pvalues")
print(summarytab, include.rownames = TRUE)
```

	ANCOVA	Stratified perm	LM perm	Residual perm
heart_freq	0.035	0.006	0.080	0.082
regurg_freq	0.136	0.118	0.280	0.220
dysp_freq	0.565	0.948	0.616	0.592
daily_heart	0.032	0.004	0.056	0.068
daily_regurg	0.142	0.174	0.286	0.246
daily_hrdq	0.043	0.012	0.088	0.098
daily_dysp	0.582	0.810	0.756	0.722

Table 3: Comparison of p-values from four tests, for each continuous endpoint.

```
print(summarytab, include.rownames = TRUE, type = "latex", file = "../ms/fig/results_pvalue_summary.tex
```