



## preference: An R Package for Two-Stage Clinical Trial Design Accounting for Patient Preference

Briana Cameron  
23andMe

Michael J. Kane  
Yale University

Denise Esserman  
Yale University

---

### Abstract

The consideration of a patient's treatment preference may be essential in determining how a patient will respond to a particular treatment. While traditional clinical trials are unable to capture these effects, the two-stage randomized preference design provides an important tool for researchers seeking to understand the role of patient preferences. In addition to the treatment effect, these designs seek to estimate the role of preferences through testing of selection and preference effects. The R package **preference** facilitates the use of two-stage clinical trials by providing the necessary tools to design and analyze these studies. To aid in the design, functions are provided to estimate the required sample size and to estimate the study power when a sample size is fixed. In addition, analysis functions are provided to determine the significance of each effect using either raw data or summary statistics. The package is able to incorporate either an unstratified or stratified preference design. The functionality of the package is demonstrated using data from a study evaluating two management methods in women found to have an atypical Pap smear.

*Keywords:* two-stage clinical trials, preference, sample size.

---

## 1. Introduction

In a traditional clinical trial setting, investigators are interested in determining the effect of a treatment on a specified outcome. The conventional clinical trial design involves the randomization of all patients to one of multiple treatment arms, which allows for the estimation of this *treatment effect*. In addition to the treatment effect, the outcome may also be influenced by a patient's preference for one of the available treatment options, especially in situations where the patient is not blinded to treatment assignment (e.g., surgical versus medical intervention). Rucker (1989) proposed a two-stage trial design in order to address this issue of patient preference. In the first stage, patients are randomized to either a choice

or a random arm. In the second stage, patients in the choice arm are allowed to select their preferred treatment, while patients in the random arm will undergo a second randomization to one of the available treatment options. In addition to the treatment effect, the two-stage design allows the estimation of two additional effects, the selection effect and the preference effect. The *selection effect* refers to the additional effect of a patient’s preferred treatment on the outcome, while the *preference effect* refers to the influence on the outcome of whether or not a patient actually receives his/her preferred treatment.

In 2016, the two-stage randomized clinical trial design was extended to allow for the inclusion of stratification factors (Cameron and Esserman 2018). The stratified version should be employed if a patient’s treatment preference is expected to be closely related to a certain measurable covariate. In this design, patients are stratified according to a specified stratification variable prior to being assigned a treatment, either through choice or randomization. For example, if older patients may be more likely to choose a medical intervention over a surgical intervention, patients would be stratified by age before randomization.

This paper introduces the R (R Core Team 2020) package **preference** for the design and analysis of two-stage (or doubly) randomized clinical trials for both the original Rucker (1989) version (unstratified) and the stratified design by Cameron and Esserman (2018). This package facilitates the implementation of the two-stage randomized design by providing the necessary sample size estimation and analysis tools for clinicians and statisticians seeking to disentangle the roles of patient preference in clinical trials. The package is available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/package=preference> (Cameron 2020).

The **preference** package fits into the general category of the “Design and Analysis of Clinical Trials” (Zhang and Zhang 2020) and focuses on Phase III trials. It is the first package to implement both unstratified and stratified clinical trial designs incorporating patient preference. The interface is simple and will readily fit into clinical trial framework packages.

## 2. The two-stage randomized trial

### 2.1. Trial design methodology

The general design, originally proposed by Rucker (1989), is illustrated in Figure 1. In this design, patients are initially randomized to either the choice or the random group. Those in the choice arm are then allowed to choose from the available treatment options, possibly after a shared decision process with their physician or family; patients in the random group are assigned a treatment through a second randomization. We assume that all patients have a preferred treatment (i.e., no undecided patients). For simplicity, we present the case where there are only two possible treatment choices.

Let  $N$  represent the total number of patients enrolled in the study and  $\theta$  represent the proportion of these patients initially randomized to the choice arm. Of note, work has been done to estimate  $\theta$  and determine the optimal allocation of participants into the choice and random arms (Walter, Turner, Macaskill, McCaffery, and Irwig 2012). We use  $\phi$  to represent the proportion of patients who prefer treatment A; because we assume all patients have a preference, we let  $1 - \phi$  represent the proportion of patients preferring treatment B.

The notations for the mean and standard deviation (SD) of each group are shown in Figure 1.

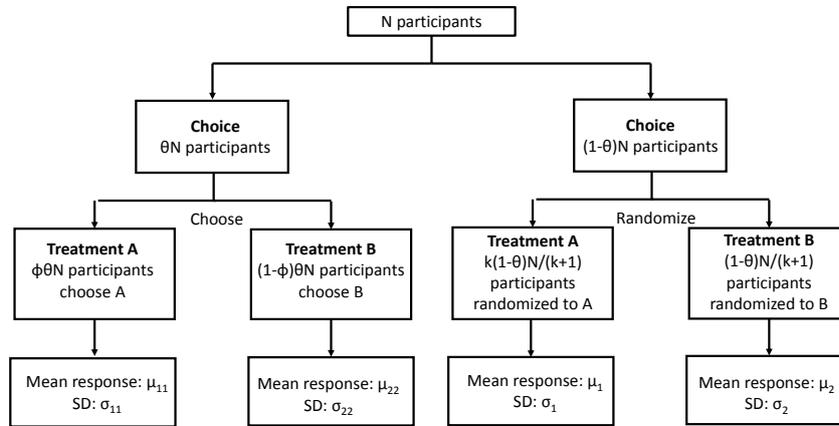


Figure 1: The flow graph for the two-stage randomized preference trial.

We use  $\mu_i$  and  $\sigma_i$  to represent the mean response and SD of patients randomized to treatment  $i$ .  $\mu_{ii}$  and  $\sigma_{ii}$  are used to represent the mean responses and standard deviations of patients choosing treatment  $i$ . For the aspects of design discussed below, we assume  $\sigma_1 = \sigma_2 = \sigma_{11} = \sigma_{22} = \sigma$ .

In cases where a treatment may not be equally preferred across the study population, the above design can be extended to include important stratification variables (Cameron and Esserman 2018). The notation is similar to that of the unstratified case, with the addition of  $\xi_l$  used to denote the proportion of patients in stratum  $l$  ( $\sum_{l=1}^s \xi_l = 1$ ).

## 2.2. Package overview

The design and analysis functions in **preference** assume a continuous, approximately normally distributed outcome measure. Analyses may be conducted on raw data or summary data. The raw data must contain the outcome variable of interest, treatment indicator, and arm (i.e., choice or random). If the design is stratified, there must also be a stratum indicator. The summary data must contain the mean and SD of the outcome, and sample size for each arm by treatment combination. These should be broken down further by stratum if the design is stratified. This package includes methods to compute treatment effects, standard error estimates, test statistics,  $p$  values and  $(1 - \alpha)$  confidence intervals for the preference, selection, and treatment effects. In the functions `preference()` and `fit_preference()`, the study results are computed from provided raw data. An alternative analysis function is provided with `fit_preference_summary()`, which computes statistics using summary data.

### *Preference trial data structure*

To aid in the design of the two-stage randomized trial, functions are provided to determine the number of participants required to detect a treatment effect, preference effect, and/or selection effect at a specified Type I and Type II error level (Turner, Walter, Macaskill, McCaffery, and Irwig 2014; Cameron and Esserman 2018). A set of preference trials is parameterized as

Parameter	Column name	Type
Preference sample size	<code>pref_ss</code>	numeric
Preference effect	<code>pref_effect</code>	numeric
Selection sample size	<code>selection_ss</code>	numeric
Selection effect	<code>selection_effect</code>	numeric
Treatment sample size	<code>treatment_ss</code>	numeric
Treatment effect	<code>treatment_effect</code>	numeric
Significance	<code>alpha</code>	numeric
Treatment ratio in random arm	<code>k</code>	numeric
Patient proportion preferring treatment 1	<code>pref_prop</code>	list of vectors with 1 numeric per stratum
Patient proportion in choice arm	<code>choice_prop</code>	numeric
Patient proportion in each stratum	<code>stratum_prop</code>	list of vectors with 1 numeric per stratum
Variance estimate of outcome of interest	<code>sigma2</code>	list of vectors with 1 numeric per stratum
Treatment power	<code>treatment_power</code>	numeric
Selection power	<code>selection_power</code>	numeric
Preference power	<code>pref_power</code>	numeric

Table 1: Columns and types of ‘`preference.trial`’ objects.

a ‘`data.frame`’ whose class attributes are ‘`preference.trial`’ and ‘`data.frame`’ with each row corresponding to an individual trial. Each trial is parameterized by the parameters shown in Table 1.

Functions are provided to construct ‘`preference.trial`’ objects based on three common scenarios. Each of these scenarios assume the effect sizes (`pref_effect`, `selection_effect`, and `treatment_effect`), the proportions (`pref_prop`, `choice_prop`, `stratum_prop`), and distributional parameters (`k`, `alpha`, `sigma2`) are known. When the sample size for each arm is known the `preference.trial()` function may be used to specify a set of trials. For the more common cases, when the maximum sample size of any arm is known, the `pt_from_ss()` function is used, and when the power must be above a threshold for each arm the `pt_from_power()` function is used. Each of these functions may take vector arguments to create multiple trials at once. When the lengths do not agree, value recycling occurs so that the number of rows is equal to the argument vector with the longest length. As a result, sets of trials can be quickly created based on many different configurations allowing the user to investigate the trial characteristics and select the best trial from a large set of possible trials.

### 3. Case study: IMAP

In the Improving Management of Mildly Abnormal Pap Smears (IMAP) study, a two-stage randomized trial design was used to evaluate psychosocial outcomes in women found to have

atypical squamous cells of undetermined significance (ASCUS) with a Pap smear (McCaffery, Irwig, Turner, Chan, Macaskill, Lewicka, Clarke, Weisberg, and Barratt 2010; McCaffery, Turner, Macaskill, Walter, Chan, and Irwig 2011). In this study, the investigators sought to determine whether women given an informed choice between two systems of management for Pap smear abnormalities showed improvements in psychosocial outcomes, satisfaction, and health-related quality of life (QOL) over women given no choice. The two arms were: (1) standard management, i.e., repeated Pap testing every 6 months to monitor for more serious abnormalities; and (2) use a human papillomavirus (HPV) triage approach, which tests for the presence of HPV types that are associated with an increased risk of cervical cancer. The original trial enrolled 330 women, 314 were randomized (16 did not meet eligibility criteria), and 235 met criteria for inclusion in the analysis. A decision aid was used to support the choice of women initially randomized to the choice arm of the study. The primary outcome for this example is health related QOL measured with the mental component score (MCS) of the 36 question short form health survey (SF36; Ware Jr and Gandek 1994). While the original trial did not stratify on any factors, for demonstration purposes, we stratified according to the six item abbreviated state trait anxiety inventory (STAI; Marteau and Bekker 1992). The STAI score is a measure of a person's anxiety, with higher STAI scores indicating higher levels of anxiety. A cutoff of 10 was the threshold for stratification (low versus high). We were kindly provided with summary data for this trial by the original investigators. We used this summary data to simulate a data set representative of the 208 individuals with stratification information from the original trial. The simulated data raw set (`imap`), as well as summarized versions of this data set (`imap_summary` and `imap_stratified_summary`) are available with the package (see Section 3.3).

### 3.1. Sample size determination

Suppose we wish to estimate the required sample size to have sufficient power for detecting treatment, preference, and selection effects for the SF36 MCS outcome in the IMAP study. First, we need to define the size of the preference, selection, and treatment effects we would like to detect. If the effect sizes are known, these effect sizes are used directly in the `pt_from_power` function. If not, the function `effects_from_means()` can be used to compute these effects from the means of each treatment group in both the choice and random arm.

In the IMAP study, we may expect the HCV triage method to result in only a small increase in SF36. That is, we expect women randomized to the HCV triage method to have a mean SF36 score of 46 ( $\mu_1$ ), while women randomized to repeated Pap screening to have a mean SF36 score of 45 ( $\mu_2$ ). In the choice arm, we expect the mean SF36 score to increase by 4 points in both management groups. Therefore, we assume that women choosing the HCV triage method to have a mean score of 50 ( $\mu_{11}$ ) and women choosing the repeated Pap smear management option to have a mean SF36 of 49 ( $\mu_{22}$ ). In addition, we must specify the proportion of women preferring the HPV triage management method. Because management with an HPV triage approach is generally a speedier option, we expect 70% ( $\phi$ ) of women to prefer this approach, as compared to the 30% of women who would prefer management with repeated Pap smears. With these assumptions, we can calculate the preference, selection and treatment effects.

```
R> effects_from_means(mu1 = 46, mu2 = 45, mu11 = 50, mu22 = 49, phi = 0.7)
```

```

$treatment
[1] 1

$selection
[1] 3.809524

$preference
[1] 9.52381

```

Therefore, with the mean SF36 scores assumed above, we need to design a study to detect a treatment effect of 1, selection effect of 3.81, and a preference effect of 9.52. To determine the necessary sample size, we must also specify the variance of the responses. In this example, we assume a variance of 5 ( $\sigma^2$ ) for the SF36 outcome. Then, the sample sizes to detect the above effects with 80% power and a Type I error of 0.05 is given by the `pt_from_power()` function.

```

R> imap_trial <- pt_from_power(power = 0.8, pref_effect = 9.52,
+   selection_effect = 3.81, treatment_effect = 1, sigma2 = 5,
+   pref_prop = 0.7)
R> class(imap_trial)

[1] "preference.trial" "data.frame"

R> imap_trial

  pref_ss pref_effect selection_ss selection_effect treatment_ss
1      23      9.52         224           3.81           314
  treatment_effect alpha k pref_prop choice_prop stratum_prop sigma2
1                1 0.05 1      0.7         0.5           1      5
  treatment_power selection_power pref_power
1             0.8             0.8         0.8

```

The function returns an object of class ‘`preference.trial`’ and ‘`data.frame`’. The trial includes three sample sizes, `pref_ss` – the preference effect sample size, `selection_ss` – the selection effect sample size, and `treatment_ss` – the treatment effect sample size. In order to achieve at least 80% power for each of the effects, the largest sample size should be selected. In this example, 314 patients are needed to detect the specified effects with at least 80% power each with a Type I error rate of 5%. If all three of these effects are of primary interest, the Type I error rate should be adjusted to account for multiple testing (e.g., Bonferroni correction).

Since it is common during the design phase to explore the effect on the sample size by varying the power, a vector of power constraints can be specified returning a set of trial designs. Additionally, a `plot()` function has been implemented to visualize the power vs. sample size relationship, shown in Figure 2.

```

R> imap_trials <- pt_from_power(power = seq(0.7, 0.9, by = 0.01),
+   pref_effect = 9.52, selection_effect = 3.81, treatment_effect = 1,
+   sigma2 = 5, pref_prop = 0.7, alpha = 0.0167)
R> plot(imap_trials)

```

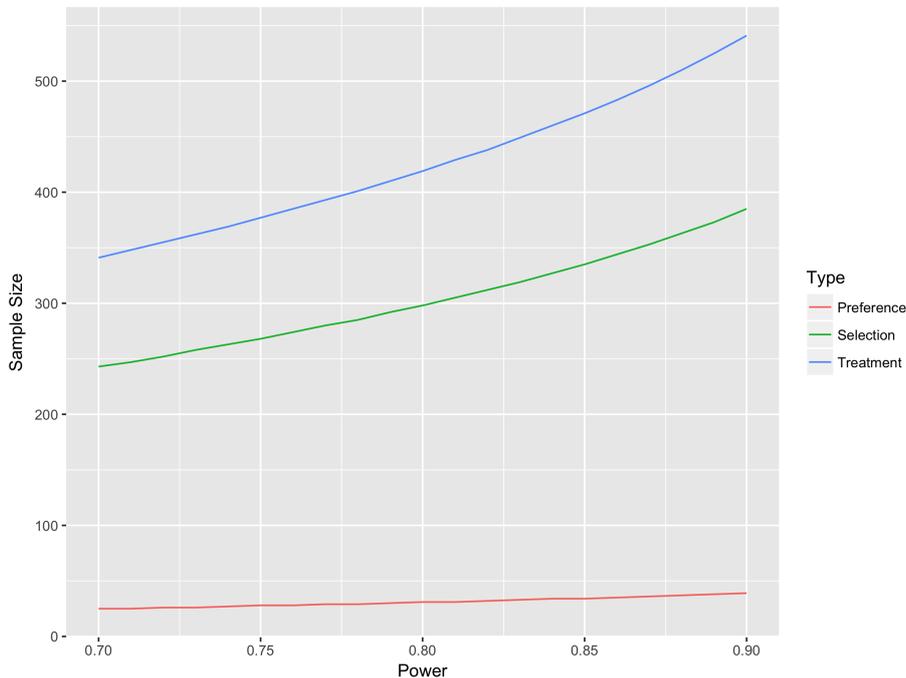


Figure 2: The effect of varying the power on sample size in the IMAP trial design.

### *Sample size determination under stratification*

We may also be interested in stratifying for STAI score (low vs. high) in our study. Because women with low STAI scores have decreased anxiety, we expect women in this stratum to have improved SF36 compared to women in the high STAI stratum. Therefore, we expect women randomized to HCV triage with low STAI scores to have a mean SF36 of 52, while women randomized to HCV triage with high STAI scores to have a mean SF36 of 42. In addition, we assume that women randomized to repeated Pap screening will have a mean SF36 of 50 in the low STAI stratum and a mean SF36 of 42 in the high STAI group.

In the choice arm, we expect both strata to have a small increase in SF36 across both management systems. For the low STAI group, we expect a 2 point increase in mean SF36 for both management groups. In the high STAI group, we also expect a 2 point increase in mean SF36 for women choosing repeated Pap screening. Because the HCV triage method is a speedier approach, we expect a slightly greater increase in SF36 for women in the high STAI group choosing this method; we assume a mean SF36 score of 46 for this group.

Finally, we assume that both management systems will be equally preferable to women in the low STAI stratum ( $\phi_1 = 0.5$ ). Because women with high STAI score have increased anxiety, we may expect a greater proportion of these women to prefer the HPV triage management method, which is the speedier option. Therefore, we assume that 70% of the women in the high STAI stratum will prefer HPV triage, and only 30% will prefer repeated Pap screening ( $\phi_2 = 0.7$ ). Furthermore, we assume that 40% ( $\xi_1$ ) of women will have low STAI scores and 60% ( $\xi_2$ ) of women will be in the high STAI score stratum. Based on these assumptions, we can calculate the preference, selection and treatment effects we wish to test in our study.

```
R> effects_from_means(mu1 = c(52, 42), mu2 = c(50, 42), mu11 = c(54, 46),
+   mu22 = c(52, 44), phi = c(0.5, 0.7), nstrata = 2, xi = c(0.4, 0.6))
```

```
$treatment
```

```
[1] 0.8
```

```
$selection
```

```
[1] 3.142857
```

```
$preference
```

```
[1] 6.457143
```

With the assumptions listed above, we will need to design a study that can detect an overall treatment effect of 0.8, selection effect of 3.14, and preference effect of 6.46. We also need to specify the variance of the responses. We assume a variance of 3 for the SF36 outcome in the low STAI stratum and a variance of 4 in the high STAI stratum. With 80% power and a Type I error rate of 0.05, we can use the `pt_from_power` function to find the sample size needed to detect the above effect.

```
R> pt_from_power(power = 0.8, pref_effect = 6.46, selection_effect = 3.14,
+   treatment_effect = 0.8, sigma2 = c(3, 4), pref_prop = c(0.5, 0.7),
+   stratum_prop = c(0.4, 0.6))
```

```
  pref_ss pref_effect selection_ss selection_effect treatment_ss
1      25      6.46      157           3.14          354
  treatment_effect alpha k pref_prop choice_prop stratum_prop sigma2
1           0.8 0.05 1 0.5, 0.7           0.5 0.4, 0.6 3, 4
  treatment_power selection_power pref_power
1           0.8           0.8           0.8
```

In the two-stage randomized design stratified by STAI score, we would need an overall sample size of 354 patients.

### 3.2. Power determination

Alternatively, we may have a pre-specified number of women to be enrolled (e.g., 208), or a range of sample sizes to explore (e.g., 200 to 500 by 50) and instead want to know the power of our study to detect each effect. If we assume the same preference, treatment, and selection effects as above (Section 3.1) and make the same assumptions regarding the preference rate and response variance, we can use `pt_from_ss()` to estimate the study power, assuming a Type I error of 0.0167. The calculation below (row 6) shows with a sample size of 450, we would have over 80% power for each of the effects, accounting for multiple testing with the Bonferroni correction. As with the sample size determination, a visualization of the trials can be created with the `plot()` function, whose output is shown in Figure 3.

```
R> pt_trials <- pt_from_ss(ss = seq(200, 500, by = 50), pref_effect = 9.52,
+   selection_effect = 3.81, treatment_effect = 1, sigma2 = 5,
+   pref_prop = 0.7, alpha = 0.0167)
R> pt_trials
```

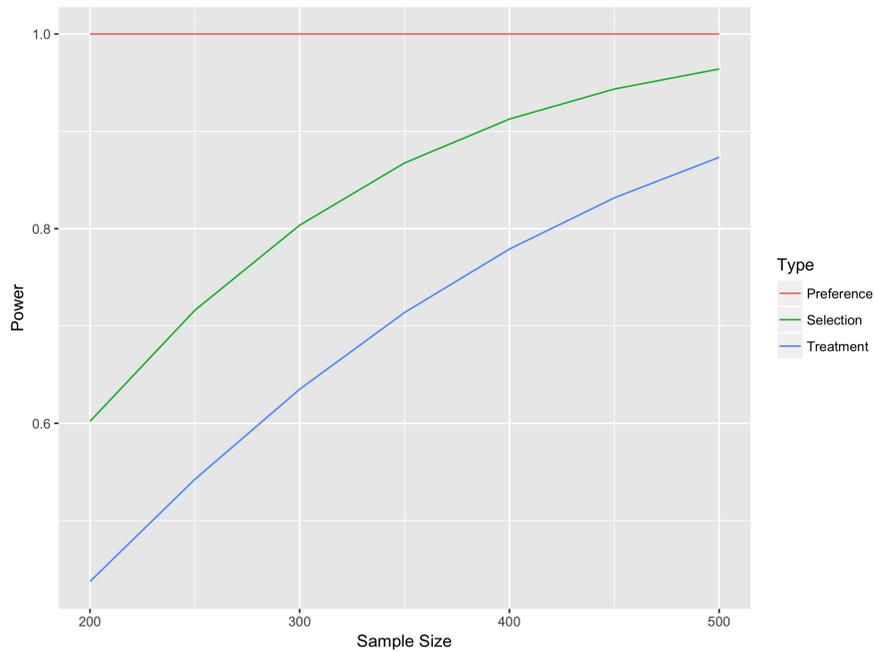


Figure 3: The effect of varying the power on sample size in the IMAP trial design.

```

pref_ss pref_effect selection_ss selection_effect treatment_ss
1      200      9.52      200      3.81      200
2      250      9.52      250      3.81      250
3      300      9.52      300      3.81      300
4      350      9.52      350      3.81      350
5      400      9.52      400      3.81      400
6      450      9.52      450      3.81      450
7      500      9.52      500      3.81      500
treatment_effect  alpha k pref_prop choice_prop stratum_prop sigma2
1                1 0.0167 1      0.7      0.5      1      5
2                1 0.0167 1      0.7      0.5      1      5
3                1 0.0167 1      0.7      0.5      1      5
4                1 0.0167 1      0.7      0.5      1      5
5                1 0.0167 1      0.7      0.5      1      5
6                1 0.0167 1      0.7      0.5      1      5
7                1 0.0167 1      0.7      0.5      1      5
treatment_power  selection_power  pref_power
1      0.4375519      0.6020127      1
2      0.5425076      0.7161912      1
3      0.6350903      0.8035972      1
4      0.7138927      0.8675243      1
5      0.7790625      0.9126054      1
6      0.8316875      0.9434559      1
7      0.8733326      0.9640403      1

```

```
R> plot(pt_trials)
```

*Power determination under stratification*

We may also wish to know our study power for a two-stage randomized trial stratified by STAI score with 208 patients. We will continue to assume that we wish to detect a treatment effect of 0.8, selection effect of 3.14, and preference effect of 6.46 and retain the assumptions above regarding the variance and preference rate within each stratum.

```
R> pt_from_ss(ss = 208, pref_effect = 6.46, selection_effect = 3.14,
+   treatment_effect = 0.8, sigma2 = c(3, 4), pref_prop = c(0.5, 0.7),
+   stratum_prop = c(0.4, 0.6), alpha = 0.0167)

  pref_ss pref_effect selection_ss selection_effect treatment_ss
1     208      6.46         208           3.14         208
  treatment_effect  alpha k pref_prop choice_prop stratum_prop sigma2
1           0.8 0.0167 1  0.5, 0.7           0.5     0.4, 0.6    3, 4
  treatment_power selection_power pref_power
1     0.4038821         0.798997         1
```

Therefore, a stratified design with an overall sample size of 208 will have 40% power to detect a treatment effect of 0.8, 100% power to detect a preference effect of 6.46, and 80% power to detect a selection effect of 3.14 at an overall Type I error of 5% for the three tests.

**3.3. Analyzing two-stage clinical trial data**

Raw data for the primary outcome, SF36 MCS (`outcome`), is provided in the `imap` data frame within the `preference` package.

```
R> head(imap)

  outcome treatment  arm strata
1 53.15752      HPV choice     1
2 56.15654      HPV choice     1
3 46.50584      HPV choice     1
4 48.39339      HPV choice     1
5 56.58280      HPV choice     1
6 49.21711      HPV choice     1
```

The treatment (HPV triage, "HPV", or repeated Pap screening, "Pap") is indicated in the `treatment` column; the study arm is either the character value "choice" (denoting the choice arm) or "random" (denoting the random arm) in the `arm` column; and the `strata` column contains the stratum indicator: `strata = 1` indicates the low STAI group and `strata = 2` indicates the high STAI group.

Analyses can be conducted using one of the three functions: `preference`, `fit_preference`, or `fit_preference_summary`. It is important to note, all three functions produce identical output; they only vary in how the data is inputted.

*Analyzing preference data with the formula interface*

`preference` allows the user to enter the model description as a formula. The functionality is similar to that of `lm`. The user is required to provide the formula and the data frame and

has the option of specifying the Type I error rate (`alpha`), which is 0.05 by default. `form` requires `outcome`, `treatment`, and `arm` (`strata` is only required if the design is stratified).

Unstratified call with default `alpha`:

```
R> pt_model <- preference(form = outcome ~ treatment:arm, data = imap)
```

Stratified call with `alpha` set at 0.1:

```
R> preference(form = outcome ~ treatment:arm | strata, data = imap,
+   alpha = 0.1)
```

```
$alpha
```

```
[1] 0.1
```

```
$unstratified_statistics
```

	[,1]	[,2]
pref_effect	2.70236332	0.701902191
pref_SE	2.96384052	4.762481400
pref_test	0.91177758	0.147381613
pref_pval	0.36188582	0.882830818
pref_LB	-2.17272050	-7.131682613
pref_UB	7.57744715	8.535486995
sel_effect	4.13557110	-11.745812237
sel_SE	2.91480874	4.343106954
sel_test	1.41881388	-2.704472250
sel_pval	0.15595328	0.006841298
sel_LB	-0.65886263	-18.889587464
sel_UB	8.93000483	-4.602037011
treat_effect	-2.85324493	3.865118139
treat_SE	1.64298661	2.170179785
treat_test	-1.73662092	1.781012876
treat_pval	0.08875226	0.078583251
treat_LB	-5.55571742	0.295490049
treat_UB	-0.15077244	7.434746228

```
$overall_statistics
```

	overall_pref_effect	overall_pref_SE	overall_pref_test	overall_pref_pval
1	1.529016	3.050353	0.5012587	0.6161891
	overall_pref_LB	overall_pref_UB	overall_sel_effect	overall_sel_SE
1	-3.488368	6.5464	-5.179471	2.818095
	overall_sel_test	overall_sel_pval	overall_sel_LB	overall_sel_UB
1	-1.837934	0.06607217	-9.814824	-0.544118
	overall_treat_effect	overall_treat_SE	overall_treat_test	overall_treat_pval
1	1.087333	1.442818	0.7536179	0.4510787
	overall_treat_LB	overall_treat_UB		
1	-1.285891	3.460558		

```
attr(,"class")
[1] "preference.fit"
```

The output is provided as a list: `$alpha` is the Type I error rate used for the confidence intervals; `$unstratified_statistics` are the stratum specific statistics with each column corresponding to a stratum; `$overall_statistics` are the statistics combined across the strata. (note that for the unstratified case `$unstratified_statistics` and `$overall_statistics` would be identical.) Both `$unstratified_statistics` and `$overall_statistics` contain the effect estimate, standard error (SE), test statistic,  $p$  value, and  $(1 - \alpha)100\%$  upper and lower bounds for each of the three effects (preference, `pref`; selection, `sel`; treatment, `treat`). For this example, none of the overall effects are statistically significant at a Type I error rate of 5%, although the overall selection effect is close to reaching statistical significance with a  $p$  value of 0.07.

#### *Analyzing data with the low-level interface*

It is also possible to specify the data elements directly with `fit_preference`. The user is required to provide the `outcome`, `arm` (the character that indicates if the arm is random, "random", or choice, "choice"), and `treatment`. If the design is stratified, then `strata`, an indicator of stratum is required. The default Type I error is 0.05, and can be changed by varying `alpha`, which should be a value between 0 and 1. Analyzing unstratified data is illustrated below using the `imap` data set.

```
R> out1 <- fit_preference(outcome = imap$outcome, arm = imap$arm,
+   treatment = imap$treatment)
```

The stratified version of this example is as follows.

```
R> out2 <- fit_preference(outcome = imap$outcome, arm = imap$arm,
+   treatment = imap$treatment, strata = imap$strata, alpha = 0.1)
```

#### *Analyzing trial summary data*

It may be that the raw data is unavailable and the user only has access to the summary data. `imap_summary` and `imap_stratified_summary` contain the summary statistics for the unstratified and stratified `imap` data set.

```
R> imap_summary
```

	mean	sd	n	treatment	arm
1	47.451	9.172	49	HPV	choice
2	47.696	9.736	74	HPV	random
3	51.489	4.321	21	Pap	choice
4	45.781	10.029	64	Pap	random

```
R> imap_stratified_summary
```

	mean	sd	n	treatment	arm	strata
1	53.175	4.517	24	HPV choice		1
2	41.957	9.192	25	HPV choice		2
3	53.005	5.080	10	Pap choice		1
4	50.111	3.120	11	Pap choice		2
5	51.163	6.399	30	HPV random		1
6	45.332	10.917	44	HPV random		2
7	54.017	5.419	22	Pap random		1
8	41.466	9.167	42	Pap random		2

The `mean` and `sd` columns indicate the mean and standard deviation of the SF36 MCS for each study group. As with the raw data, `treatment` indicates the treatment, `arm` is a logical denoting the first stage randomization to choice or random, and in `imap_stratified_summary`, `strata` indicates the stratum.

The `fit_preference_summary` function can be used to analyze the summary data by inputting the mean, variance, and sample size of each group.

```
R> out3 <- fit_preference_summary(x1mean = 47.451, x1var = 9.172^2, m1 = 49,
+   x2mean = 51.489, x2var = 4.321^2, m2 = 21, y1mean = 47.696,
+   y1var = 9.736^2, n1 = 74, y2mean = 45.781, y2var = 10.029^2, n2 = 64)
```

In a stratified design, the mean, variance, and sample size of each stratum must be included along with the number of strata (`nstrata`) and the proportion of patients within each stratum (`xi`). `alpha` may be varied from the default of 0.05 for both the stratified and unstratified versions.

```
R> attach(imap_stratified_summary)
R> x1mean <- c(mean[1], mean[2])
R> x1var <- c(sd[1]^2, sd[2]^2)
R> m1 <- c(n[1], n[2])
R> x2mean <- c(mean[3], mean[4])
R> x2var <- c(sd[3]^2, sd[4]^2)
R> m2 <- c(n[3], n[4])
R> y1mean <- c(mean[5], mean[6])
R> y1var <- c(sd[5]^2, sd[6]^2)
R> n1 <- c(n[5], n[6])
R> y2mean <- c(mean[7], mean[8])
R> y2var <- c(sd[7]^2, sd[8]^2)
R> n2 <- c(n[7], n[8])
R> xi_1 <- sum(n[1], n[3], n[5], n[7])/sum(n)
R> xi <- c(xi_1, (1 - xi_1))
R> detach()
R> out4 <- fit_preference_summary(x1mean = x1mean, x1var = x1var, m1 = m1,
+   x2mean = x2mean, x2var = x2var, m2 = m2, y1mean = y1mean,
+   y1var = y1var, n1 = n1, y2mean = y2mean, y2var = y2var, n2 = n2,
+   xi = xi, nstrata = 2, alpha = 0.1)
```

### 3.4. Optimal distribution between choice and random arms

In the examples presented above, we have assumed the choice proportion (`choice_prop`), the proportion of individuals that are distributed to the choice arm, to be 0.5 (equal distribution). In general, however, the choice proportion is not a fixed parameter and can be specified according to the main goals of the study. If a desired choice proportion is known in advance, the user can directly set this argument in the function calls above. If the desired choice proportion is not known, the function `optimal_proportion()` can be used to estimate the optimal value given the goals of a particular study (Walter *et al.* 2012).

To use `optimal_proportion()`, a set of weights must be defined that indicate the relative importance of estimating each of the three effects (`w_sel`, `w_pref`, `w_treat`) with the constraint  $w_{\text{sel}} + w_{\text{pref}} + w_{\text{treat}} = 1$ . For example, if we are equally interested in estimating the selection, preference, and treatment effects, we would assign a weight of  $\frac{1}{3}$  to each effect (i.e.,  $w_{\text{sel}} = w_{\text{pref}} = w_{\text{treat}} = 0.33$ ). If we want to determine the optimal choice proportion for the unstratified IMAP example above to detect a preference effect of 9.52 and a selection effect of 3.81 (assuming 70% of women prefer the HCV triage method and a variance of 5), we can use the following function call.

```
R> optimal_proportion(w_sel = 1/3, w_pref = 1/3, w_treat = 1/3, sigma2 = 5,
+   phi = 0.7, delta_pi = 9.52, delta_nu = 3.81)
```

```
[1] 0.6399756
```

Therefore, if we are equally interested in the estimation of the preference, selection, and treatment effects, we should allocate 64% of patients to the choice arm in the initial randomization stage.

Note, this function is only available for the unstratified design at this time, although work is currently underway to expand this to the stratified case.

## 4. Summary

The two-stage randomized clinical trial design is an important tool for researchers seeking to understand the impact of treatment preference on outcome. In cases where equivalence has been demonstrated between two treatments, patient preferences may be an important consideration for health care professionals seeking to provide the best possible treatment for a particular patient. Estimation of the effect of treatment preference will also be of particular interest when blinding is not feasible and when treatment success is closely tied to a patient's adherence.

This paper introduces the R package **preference** for the design and analysis of these trials. The IMAP study discussed above is used to demonstrate the main functionality of the package. To aid in the design of the two-stage randomized trial, sample size formulas are provided to estimate the required sample size to detect a given preference, selection, and treatment effect. In cases where the sample size is fixed, functions to compute the study power to detect each effect are included. Finally, analysis functions are provided that can calculate the test statistic,  $p$  value, effects estimates and 95% confidence intervals for each effect using either raw or summary data. These functions can be used for both the unstratified and stratified study

designs. In addition, the package provides tools for optimizing the distribution of patients in the initial randomization between choice and random arm in the unstratified case.

This package is an important contribution as it is the first available software to design two-stage clinical trials incorporating patient preference. Previous simulation studies are neither available as part of supplementary material with the paper nor publicly available from the investigators over the web. Extensions to the **preference** package are being actively developed concurrently with extensions to the methodology. The above methods assume a continuous, approximately normally distributed outcome measure; extensions to these methods are being developed to allow for the design and analysis of studies with a binary or count endpoints.

## Acknowledgments

This work was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1511-32832) and Yale’s CTSA Award (U11TR001863). We would also like to thank the IMAP team for sharing their data to demonstrate this package.

## Disclaimer

All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

## References

- Cameron B (2020). **preference**: *2-Stage Preference Trial Design and Analysis*. R package version 1.1.6, URL <https://CRAN.R-project.org/package=preference/>.
- Cameron B, Esserman DA (2018). “Sample Size and Power for a Stratified Doubly Randomized Preference Design.” *Statistical Methods in Medical Research*, **27**(7), 2168–2184. doi:10.1177/0962280216677573.
- Marteau TM, Bekker H (1992). “The Development of a Six-Item Short-Form of the State Scale of the Spielberger State-Trait Anxiety Inventory (STAI).” *British Journal of Clinical Psychology*, **31**(3), 301–306. doi:10.1111/j.2044-8260.1992.tb00997.x.
- McCaffery KJ, Irwig L, Turner R, Chan SF, Macaskill P, Lewicka M, Clarke J, Weisberg E, Barratt A (2010). “Psychosocial Outcomes of Three Triage Methods for the Management of Borderline Abnormal Cervical Smears: An Open Randomised Trial.” *BMJ*, **340**. doi:10.1136/bmj.b4491.
- McCaffery KJ, Turner R, Macaskill P, Walter SD, Chan SF, Irwig L (2011). “Determining the Impact of Informed Choice: Separating Treatment Effects from the Effects of Choice and Selection in Randomized Trials.” *Medical Decision Making*, **31**(2), 229–236. doi:10.1177/0272989x10379919.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

- Rucker G (1989). “A Two-Stage Trial Design for Testing Treatment, Self-Selection and Treatment Preference Effects.” *Statistics in Medicine*, **8**(4), 477–485. doi:10.1002/sim.4780080411.
- Turner RM, Walter SD, Macaskill P, McCaffery KJ, Irwig L (2014). “Sample Size and Power When Designing a Randomized Trial for the Estimation of Treatment, Selection, and Preference Effects.” *Medical Decision Making*, **34**(6), 711–719. doi:10.1177/0272989x14525264.
- Walter SD, Turner RM, Macaskill P, McCaffery KJ, Irwig L (2012). “Optimal Allocation of Participants for the Estimation of Selection, Preference and Treatment Effects in the Two-Stage Randomised Trial Design.” *Statistics in Medicine*, **31**(13), 1307–1322. doi:10.1002/sim.4486.
- Ware Jr JE, Gandek B (1994). “The SF-36 Health Survey: Development and Use in Mental Health Research and the IQOLA Project.” *International Journal of Mental Health*, **23**(2), 49–73. doi:10.1080/00207411.1994.11449283.
- Zhang E, Zhang HG (2020). “CRAN Task View: Clinical Trial Design, Monitoring, and Analysis.” Version 2020-01-12, URL <https://CRAN.R-project.org/view=ClinicalTrials>.

**Affiliation:**

Briana Cameron  
23andMe  
San Francisco, CA, United States of America  
E-mail: [bri.cameron11@gmail.com](mailto:bri.cameron11@gmail.com)

Michael J. Kane and Denise Esserman  
School of Public Health  
Biostatistics Department  
Yale University  
New Haven, CT, United States of America  
E-mail: [michael.kane@yale.edu](mailto:michael.kane@yale.edu), [denise.esserman@yale.edu](mailto:denise.esserman@yale.edu)