

# A permutation-based funnel plot for subgroup analysis

Carlo Veltri (✉ [carlo.a.g.veltri@gmail.com](mailto:carlo.a.g.veltri@gmail.com))  
<https://orcid.org/0009-0006-8726-3293>



---

## Short Report

## Keywords:

**Posted Date:** August 8th, 2023

**DOI:** <https://doi.org/10.21203/rs.3.rs-3227587/v2>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

---

# A permutation-based funnel plot for subgroup analysis

Carlo Veltri, Bodo Kirsch, Hermann Kulmann

2023-08-02

## Introduction

The purpose of clinical trials is to evaluate the safety and efficacy of a new treatment or intervention in order to assess its potential usefulness in clinical practice. However, efficacy and safety of a drug or intervention can vary largely among subgroups. Accounting for variations in treatment responses has also come the attention of regulatory authorities; according to the EMA guidelines for subgroup analysis in clinical trials (1), factors considered relevant for treatment response can be e.g. demographics, genomic factors, clinical patient characteristics or regional effects, which should be taken into account both when planning a study and when evaluating its results.

Screening for relevant treatment-covariate interactions has been kept to a minimum for a long time due to computational capacities, which have remarkably improved in the last years. However, post-hoc subgroup screening still has a mixed reputation due to accusations of cherry picking favorable results for subgroups when “rescuing” a failed study (2), e.g., by identifying subgroups with a seemingly positive outcome. As much as post-hoc subgroup analysis can be used in bad faith, it is necessary for ethical considerations, as finding relevant risks or benefits for subgroups is required both when approving a new drug or therapy and when planning further clinical trials to investigate said benefits or risks.

Muysers et al. (2) propose a new method for explorative post-hoc subgroup screening, the *Subgroup Explorer*, an interactive R Shiny app available free of charge in the R package *subscreen* on CRAN (<https://cran.r-project.org/>). The proposed method consists of (i) calculating point estimates for all subgroups in a clinical trial and (ii) interactive plotting of the estimates in an R Shiny app.

While the *Subgroup Explorer* is already in use and widely applied in randomized clinical trials as well as observational studies such as (3), clinical study teams have communicated their demand for a further extension of the visualization: the point estimates for subgroups should be accompanied by confidence intervals accounting for the relation of effect size and variance of a treatment effect in specific subgroups. In this article, one such extension is proposed and evaluated. Confidence intervals are calculated based on the inversion of permutation tests, whose popularity in subgroup analysis increased recently (4). Permutation-based testing as well as the specific approach used to construct confidence intervals in this setting are discussed in the subsequent chapters, and the method’s performance is tested on a simulation based on synthetic clinical trial data. Afterwards, strengths and limitations are discussed.

There are other methods for subgroup identification, discovery and confirmatory analysis. Some take advantage of recent advances in machine learning, such as shrinkage estimators or tree-based methods for finding subgroups that are most predictive for treatment responses. A broad overview of those methods is given in (5). Such methods should be considered complementary to the visualization approach of the *Subgroup Explorer* in identifying remarkable subgroups, and as a necessary follow up for confirmatory statistical analysis when multiplicity adjustments are applied.

The main innovation proposed can be put in three distinct considerations: (i) the inversion of permutation tests, which is rarely applied in practice; (ii) obtaining confidence intervals through such method for a larger number of point estimates, the confidence intervals only being dependent of the sample size of the subgroup and (iii) connecting these confidence intervals, turning them into a two-dimensional *confidence region* along the funnel plot.

Lastly, it should be emphasized that the resulting confidence intervals are not to be interpreted as confirmatory for any subgroups effects as they are used purely for exploratory reasons. We further use the term *confidence interval* since the estimates are based on the inversion of a statistical testing procedure, however, (and as is the case for any post-hoc subgroup analysis) results should only be regarded as exploratory and only allow researchers to screen for any cases that are of interest for further investigation - which should then make use of confirmatory statistical analyses.

## Methods

### The Subgroup Explorer

In this setting, subgroups are a combination of levels of different factors, where factors are a unique characterization of an observation (e.g., *sex*). Levels refer to a specific realization of it (e.g., *female*). All factors must be either categorical or classified continuous (e.g. BMI with levels  $<20$ ,  $[20,25]$  and  $>25$ ). The data must also contain the endpoints of interest, treatment groups and an evaluation function. Individual endpoints can, for instance, be continuous endpoint measures or time to event data (e.g. time to disease progression) and are input to the evaluation function, which will calculate a numeric summary measure, such as a proportion or difference thereof, a mean or median, or a hazard ratio.

First, all possible subgroups are calculated. For each existing subgroup, all observations falling into that category are taken as input to the evaluating function. This generates point estimates for the evaluation function for each subgroup. For example, taking sex and BMI (classified as low, medium or high) as input factors as well as some continuous endpoint in a two-arm trial for which the average treatment effect is calculated, the average treatment effect would be calculated in each subgroup  $\{female, male, female \times BMI_{low}, male \times BMI_{low}, \dots, male \times BMI_{high}\}$ , where  $\times$  indicates the interaction between the two factors.

Next, the point estimates for subgroup-specific treatment effects are visualized. For reasons of interpretability and visual capacity, subgroups are displayed up to combined factors of three levels. Each subgroup is represented by one dot, where the y-axis indicates the target variable (in most cases, the treatment effect) and the x-axis indicates the subgroup size, both as an absolute value and the relative size in percentages of the study population.

Several other features for the *Subgroup Explorer* exist, such as interactive filtering and estimation of predictive covariates by random forests. We will further limit the discussion to the main specification (i.e. plotting subgroup treatment effects on the size of the subgroup) as the other features are complementary and independent of the proposed funnel model. We recommend referring to the CRAN manual (6) and (2) for a comprehensive overview of features.

Figure 1 shows a graphical interface of the *Subgroup Explorer* with simulated treatment and endpoint data, where the subgroup-defining factors and their multivariate distribution are taken from actual RCT data in (7). The endpoint was drawn from a normal distribution with mean one in the treated group and mean zero in the control group (an average treatment effect of one). Each dot represents a single subgroup characterized by up to three factors. The overall treatment effect is represented by a horizontal line. For homogeneous treatment effects, a funnel-shaped scatter diagram around the overall treatment effect is expected, as larger subgroups have a larger effect on the overall treatment effect whereas smaller subgroups can have effects more distant to the overall treatment effect. As the simulated treatment effect was drawn independently from the subgroups, the given funnel represents a typical plot of homogenous treatment effects. The further discussion aims to answer how to estimate a funnel in which  $100 \times (1 - \alpha)$  percent of the subgroups lie within under the assumption of homogeneous treatment effects. What proceeds next is a short overview on permutation testing and its application in this case.

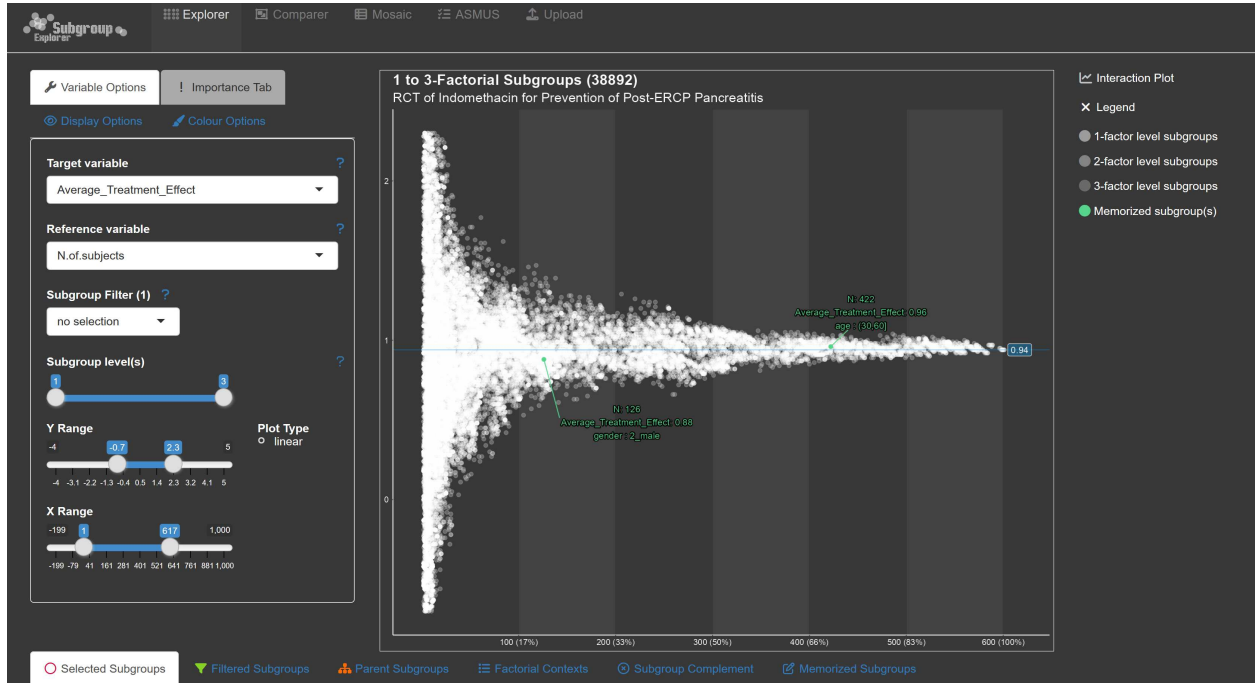


Figure 1: Typical interface of the subgroup explorer with simulated endpoint and treatment arms. The subgroups are independent of the simulated treatment and taken from an RCT to mimic a realistic distribution of subgroups given treatment effect homogeneity. In total there are 38892 subgroups, defined by up to 3 factors. Two subgroups were memorized and shown in light green: the age group of 30 to 60 year olds (422 subjects) and male patients (126).

## Permutation based testing

Permutation tests are non-parametric tests designed to test associations between samples of two random variables  $(x_i, y_i)$ . The distribution of the test statistic (e.g. mean difference) is approximated by obtaining new samples through permutations of  $X$  and  $Y$ . Either all permutations or a random subset can be used. The former approach is called the *exact* method whereas the latter approach is often referred to as the *Monte Carlo* method (often implemented by drawing without replacement, cf. (8)). P-values are obtained from the approximated null distribution of the test statistic derived from the permuted samples.

Advantages of permutation tests over parametric tests were discussed in length, cf. (9). Most noteworthy is flexibility as permutation tests can be applied to a variety of estimators, i.a. arithmetic and geometric means, percentages or ranks. For subgroup analysis in randomized clinical trials (RCTs), there are some particular considerations with respect to modelling the desired null scenario. Those considerations are discussed in more depth in (4).

RCTs can be characterized by the set  $(y_i, T_i, X_i), i = 1, \dots, n$ , where  $y_i$  is the observed outcome (endpoint) of patient  $i$ ,  $T_i$  is the treatment level and  $X_i$  are covariates of interest. Permuting all components leads to a model where all associations are removed, making the treatment effect under the null hypothesis zero. (10) suggests to permute  $y$  within levels of  $T$ , that is, to permute  $(y_i, T_i)$  jointly, which is equivalent to permuting the covariates  $X_i$  jointly and keeping  $(y_i, T_i)$  fixed (cf. (4)).

## Implementation in the Subgroup Explorer

Consider one subgroup  $A$  and its complement subgroup  $A^C$  such that  $A \cup A^C$  form the study population of a two-arm RCT. The treatment effect in subgroup  $A$  (and  $A^C$  respectively) is estimated by

$$\hat{\mu}_A = \frac{1}{n_{A_1}} \sum_{i \in A, T_i=1} y_i - \frac{1}{n_{A_0}} \sum_{i \in A, T_i=0} y_i$$

where  $n_{A_1}$  and  $n_{A_0}$  are the number of subjects in subgroup  $A$  with treatment  $\{1, 0\}$  such that  $n_{A_1} + n_{A_0} = n_A$  is the number of subjects in subgroup  $A$ . A 95%-confidence interval for  $\hat{\mu}_A$  can be constructed by the procedure:

1. Randomly shuffle (permute) values of  $(y_i, T_i)$  between  $A$  and  $A^C$  to obtain a new sample  $A'$  by drawing out of  $A \cup A^C$   $n_A$  times without replacement
2. Estimate  $\hat{\mu}_{A'}$
3. Repeat  $k = 1000$  times and obtain the 5% and 95%-quantile

As the goal is to find a two-dimensional confidence region over the whole scatter plot, representative sizes for hypothetical subgroups (referred to as *support points* furtherly) are chosen. 50 support points in total are estimated in equal steps of the square-root of the minimum and maximum subgroup size, because larger decreases in variation are expected for smaller subgroup sizes. This mimics the  $\sqrt{n}$ -convergence of many estimators (e.g. the sample mean). The support points are connected through local polynomial regression, provided by the LOESS function of the R stats package (11), with a decreased span (0.25) compared to the default (0.75), as larger variations in the confidence intervals are again expected for small sample sizes, such that confidence intervals for larger subgroups should not take those random variations too much into account.

## Results

A full replication package for the study results is available on [github.com/veltrica](https://github.com/veltrica). As in the previous chapter, a synthetic RCT was created by taking covariates from (7) and simulating a continuous average treatment effect. The analysis considers a scenario where no treatment effect heterogeneity is present, i.e. the subgroups are independent of the treatment effect. The subgroups were fixed for each simulation such that there was no variation in subgroup sizes. For 500 simulations in total, the following procedure was implemented:

- Randomly allocate half of the study population to treatment and control group
- Randomly draw the endpoint from a standard normal distribution and add +1 if the subject is in the treated group
- Calculate confidence intervals for each support point
- Save the proportion of points outside for each  $\alpha \in [1\%, 5\%, 10\%]$

The number of permutations per support point  $n_p$  was varied for  $n_p \in [1000, 2000, 5000]$  to investigate whether adding further permutations increases the accuracy of the reference funnel. The proportion of points outside of the reference funnel is considered as an analog to the false discovery rate in simulations of tests. It is calculated by the number of subgroups which are outside the estimated confidence interval (i.e. an actual support point estimate or the interpolated LOESS prediction between two support points) divided by the number of all subgroups. For the funnel to be considered “exact” in hypothesis testing terms, the proportion of points outside should be close to the desired significance level  $\alpha$ . As potentially bad small-sample confidence intervals and specifically their LOESS extrapolations could distort the results, subgroups outside the funnel were only considered at a subgroup size of at least ten percent of the study population (at least 60 patients), which accounted for around 11 000 subgroups. However, this threshold only avoids taking into account small-sample properties of the funnel; whether the effect in a subgroup given its size is large enough to be considered promising requires both statistical and clinical judgement and should be decided on a case-to-case basis.

	$n_p = 1000$	$n_p = 2000$	$n_p = 10000$
$\alpha = 1\%$	1.26% (0.08%)	1.20% (0.08%)	1.16 (0.08)
$\alpha = 5\%$	5.33% (0.19%)	5.30% (0.19%)	5.27 (0.19)
$\alpha = 10\%$	10.39% (0.27%)	10.37% (0.27%)	10.33 (0.27)

Table 1 shows the results of the performed simulation study, with the average number of points outside for different numbers of permutations  $n_p$  and for different quantiles. Standard errors of the means are given in brackets. All quantiles were slightly overestimated, but the differences decreased when the number of permutations was increased. Overall, using 1000 permutations per support point seems to be a good enough approximation.

Figure 2 displays an interface of the *Subgroup Explorer* given an estimated reference funnel. The support points’ point estimates are also shown, which can be disabled by the user. In this scenario, treatment effect heterogeneity was simulated: the treatment effect was simulated as before, but +0.5 was added if the patient was both female and in the treatment group. It is immediately visible how the search for the ‘needle in the hay stack’ is simplified, as both female and male patients (displayed as *memorized* subgroups) are outside of the 5%-reference funnel.

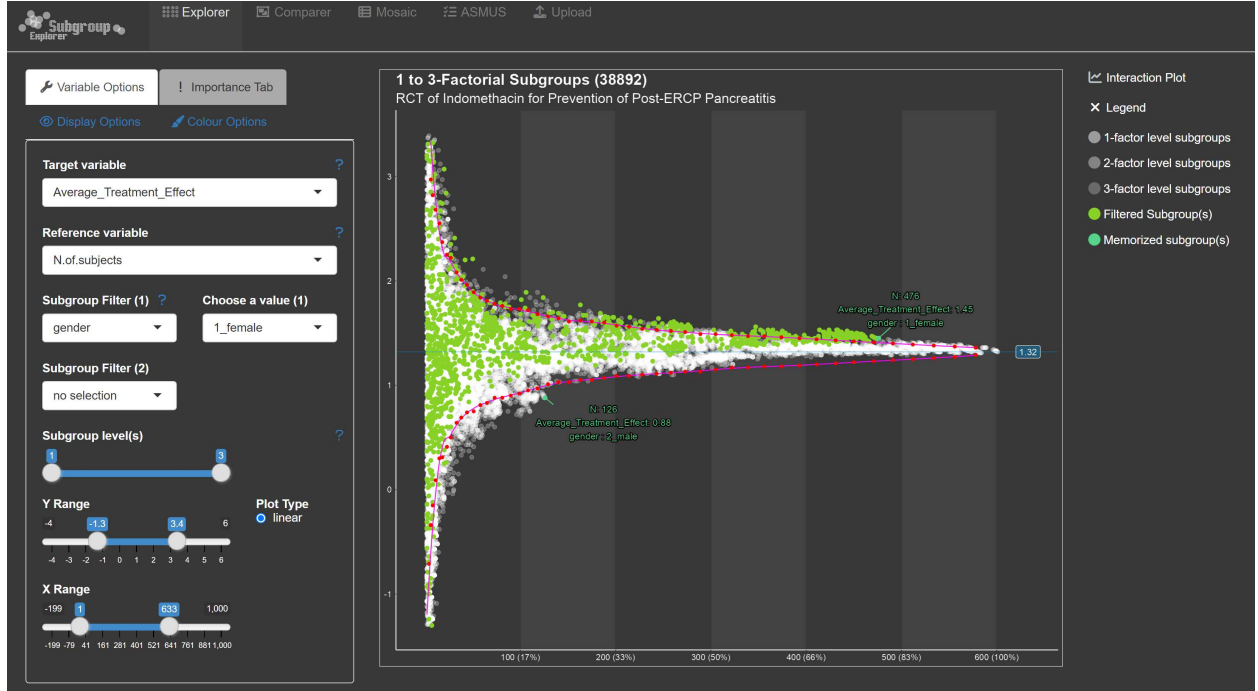


Figure 2: The Subgroup Explorer with an estimated 5%-reference funnel with 1000 permutations. The point estimates for the confidence intervals are coloured in red, and the estimated curve is coloured in magenta. All subgroups which had female as one of their factor levels are colored in green.

## Discussion

This article proposes a new method to estimate a reference funnel along the plot of the *Subgroup Explorer*. The method relies on inverting permutation tests to estimate confidence intervals for representative subgroup sizes, and estimating a smooth function to approximate the two-dimensional confidence region. This discussion will cover some strengths and limitations of the new method.

The primary strength relies in pursuing a non-parametric, easy to implement approach. This makes the method applicable to a variety of estimators, for whom the actual theoretical distribution is often difficult to obtain. Large sample approximations need not to be used. Permutation tests have the desirable property of being exact, such that the asymptotic false discovery rate is at most at the desired significance level for each level. The simulation results underline the general validity of implementing this approach.

The main limitations draw on practical considerations and limitations of subgroup analysis in general. Limitations of subgroup analysis in general are various, which are extensively discussed in (5). Two pitfalls are worth emphasizing with respect to the proposed method. Firstly, the method assumes a fixed treatment effect in the population such that only the variation of a subset around the study mean can be detected. The confidence interval will always be centered around the study treatment effect and cannot account for uncertainty with respect to the population treatment effect. This may result in false certainty about the resulting significance of a subgroup finding, and always requires both sound judgement by a statistician and scientific and clinical reasoning by the study team. Secondly, confounding variables can make the analysis more difficult. Subgroups with a high overlap (e.g. non-smokers and individuals with no history of drug abuse) will always have a similar point estimate, again requiring scientific as well as clinical reasoning. This is especially present when the confounding variable is not measured and the supposedly relevant variable is measured (e.g. only measuring smoking and not blood pressure, when the relevant subgroup is hypertension).

Practical considerations draw on computational power. We recommend to use parallel processing to reduce the computing time, which can be done by increasing the number of kernels in the *subscreencalc* function.

For computational efficiency, the number of permutations and number of support points has been kept to a minimum, which might affect the quality of the estimated confidence intervals or the overall confidence funnel respectively. For more accurate results, researchers might want to adjust these arguments in the *subscreencalc* function.

Much research can be put in the *Subgroup Explorer* and the reference funnel. Further methodological advances could, among other things, include small sample properties of funnel estimation, considering other possible x-axis and their distributions (e.g., plotting the number of events rather than sample sizes in time-to-event studies) and power analysis for detecting promising subgroups with typical funnel forms that indicate treatment effect heterogeneity.

## Conclusion

The visualization approach of Bayer’s Biostatistics Innovation Center implemented in the R package *subscreen* on CRAN has contributed to the recent discussion of subgroup analysis in clinical trials. The funnel model shows to be helpful in explorative subgroup analysis with large amounts of considered factors. By a permutation-based procedure, distributional assumptions are avoided, such that the model can be implemented with a variety of estimators applied in clinical trials. Researchers still should be aware of the pitfalls of post hoc subgroup analysis, but despite all risks, but there is an ethical responsibility to fully explore clinical trial data.

## Acknowledgements

We want to thank Steffen Jeske for helping with implementation, Dr. Franco Mendolia and Jennifer Gilbride for providing useful comments as well as the Bayer Biostatistics Innovation Center for making this research possible.

## Declaration of Conflicting Interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. EMA/CHMP. Guideline on the investigation of subgroups in confirmatory clinical trials. <https://www.ema.europa.eu/en/investigation-subgroups-confirmatory-clinical-trials-scientific-guideline#current-version-section>; 2019.
2. Muysers C, Dmitrienko A, Kulmann H, Kirsch B, Lippert S, Schmelter T, et al. A systematic approach for post hoc subgroup analyses with applications in clinical case studies. *Ther Innov Regul Sci*. 2019;
3. Muysers C, Messina F, Keil T, Roll S. A novel concept of screening for subgrouping factors for the association between socioeconomic status and respiratory allergies. *Journal of Exposure Science & Environmental Epidemiology*. 2022;
4. Foster JC, Nan B, Shen L, Kaciroti N, Taylor JMG. Permutation testing for treatment-covariate interactions and subgroup identification. *Stat Biosci*. 2016;
5. Lipkovich I, Dmitrienko A, D’Agostino RBSr. Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*. 2016;
6. Kirsch B, Lippert S, Schmelter T, Muysers C, H. K. Systematic screening of study data for subgroup effects. CRAN manual. R Foundation for Statistical Computing; 2018.
7. Higgins P. Data package for medical datasets. CRAN manual. R Foundation for Statistical Computing; 2021.



8. Berry KJ, Johnston JE, Mielke PWJ. Permutation methods. WIREs Computational Statistics. 2011;
9. Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. Clinical and Experimental Pharmacology and Physiology. 1994;
10. Potthoff RF, Peterson BL, George SL. Detecting treatment-by-centre interaction in multi-centre clinical trials. Statistics in Medicine. 2001;
11. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: <http://www.R-project.org/>