Graphical displays for subgroup analysis in clinical trials

Yi-Da Chiu[1], Nicolas Ballarini[2], Franz Koenig[2], Martin Posch[2] and Thomas Jaki[1*]

1. Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, LA1 4YF, U.K.
2. Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.

* t.jaki@lancaster.ac.uk

Abstract

Subgroup analyses are a routine part of clinical trials to investigate the effect of treatments in subsets of the population under study. The purpose of this assessment may be to ensure that there are no groups of patients for whom the treatment is harmful despite being effective in the majority of patients or to identify groups of patients that may benefit from a treatment when the overall effect is small or zero. This topic has received extensive attention in recent clinical research thanks to the development of stratified medicine and the availability of genomic and biomarker data.

Graphical approaches play a key role in subgroup analyses to visualize effect sizes of subgroups, aid identification of groups that respond differentially, and communicate the results to a wider audience. However, many existing approaches do not capture the core information and/or are prone to lead to misinterpretation of subgroup effects. In this work, we critically appraise existing visualization techniques, propose useful extensions to increase their utility and attempt to develop an effective visualization approach. The graphical techniques considered include level plots, contour plots, bar charts, Venn diagrams, tree plots, forest plots, Galbraith plots, L'Abbé plots, the subpopulation treatment effect pattern plot, alluvial plots and UpSet plots. We illustrate the methods using a dataset of a treatment for prostate cancer with survival outcome.

Keywords: Data visualisation, subgroup analysis, forest plot

1 Introduction

Investigating target populations that potentially benefit from an innovative intervention is essential in clinical trials. Even if efficacy is established in the overall population, a complete benefit/risk assessments of subgroups should be undertaken before deciding whether the treatment is administered to the whole population or certain subgroups need to be excluded. Such investigations pose a challenge because of the various issues that are needed to address. For example, enrolling patients that have rather diverse baseline characteristics for considerations, such as age, gender, race, disease severity or biomarker profiles may create a large number of subgroups. Also, the credibility level of the findings is concerned. The presence of promising results can be attributed to a small sample size or to the fact that many subgroups are searched.

Subgroup analyses as investigative measures are prospective or post-hoc in different settings of clinical trials. Their primary proposes can be to establish efficacy claim, subgroup discovery and consistency assessments across subgroups. It is therefore a broad field addressing various subgroup problems as mentioned before. Many researchers have proposed novel approaches and designs for different categories of subgroup analysis [1–3]. It has further received extensive attention in recent clinical research for the development of stratified medicine.

Graphical approaches are routinely employed in subgroup analysis, typically for describing treatment effect sizes of subgroups. Such visualisations encapsulate subgroup information and boost clinical decision making process. However, not much attention has been paid to how to make effective graphics in the subgroup analysis setting. Existing approaches still have inherent drawbacks and their use may lead to misinterpretations on subgroup effect sizes [1]. For instance, forest plots provide no insight on the overlap of different subgroups. Additionally, whether or not a subgroup confidence interval crosses the no-effect point does not necessarily imply a differential effect in the subgroup. It is therefore crucial to correctly depict effect sizes and essential subgroups information.

In addition to displaying treatment effects, several characteristics are desirable for graphical approaches as initial subgroup analysis tools. Showing sample sizes is necessary because it underpins the credibility level of promising and adverse findings within subgroups. Revealing overlap information also enables to focus on the subgroups which have a less overlap with each other in the final presentation. The ability of detecting heterogeneity for all subgroup treatment effect sizes should be considered as well. Moreover, it is expected to be available for large number of subgroups to serve as a potential hypothesis generator. These characteristics can certainly constitute sensible criteria for assessments.

In this paper we attempt to develop an effective visualization approach for subgroup analysis. Our considerations are not constrained regardless of exploratory or confirmatory settings. To facilitate the development, we focus on a clinical trial dataset of a treatment for prostate cancer. The graphical techniques considered include level plots, mosaic plots, contour plots, bar charts, Venn diagrams, tree plots, forest plots, Galbraith plots, L'Abbé plots, the subpopulation treatment effect pattern plot, alluvial plots, UpSet plots, and circle plots

The remainder of the paper is structured as follows: in Section 2 we describe the dataset we use for illustration, in Section 3 we present and exploit the graphical approaches for displaying subgroup information. Each technique is further assessed based on a set of criteria. Section 4 focuses on alternative graphical displays to further explore subgroup compositions. In Section 5 we summarise the assessments and features of all the improved approaches. Remarks on their practical utility and implications in clinical trials are made. We outline potential visualisation techniques in the end.

2 Dataset for illustration: The prostate cancer dataset

We use a prostate carcinoma dataset from a clinical trial [4] which is available on the web [5]. The data has been analyzed several times in the literature before, and [6] used it to illustrate subgroup analysis by model selection. The dataset consists of 475 subjects randomized to a control group or diethyl stilbestrol.

We are interested in identifying subgroups of patients that may benefit from the treatment. There are six variables to consider: existence of bone metastasis (bm), disease stage (3 or 4), performance (pf), history of cardiovascular events (hx), age, and weight.

As the considered endpoint is survival time in months, the graphical approaches use either the log-hazard ratios for treatment versus control or the differences in restricted mean survival time (RMST) as treatment effect measure. The latter is useful when want to depict estimates for control and treatment group separately.

3 Graphical approaches to subgroup problems

Several graphical approaches are used for visualization on certain information of subgroups from the prostate cancer dataset. Each of the graphical displays is first presented and then assessed based on a set of sensible criteria. We additionally point out other noticeable features of each display approach.

The criteria is prioritised as follows:

C1 whether the plot displays effect sizes for subgroups;

C2 whether it exhibits subgroup sample sizes;

C3 whether it shows all overlap information for subgroups;

C4 whether it serves for detecting heterogeneity in treatment effect sizes (or treatment-covariate interactions);

C5 whether it is available for the large number of total subgroups (more than 10)

NB: Two additional criteria may be: showing uncertainty or precision of the estimates and large number of subgroup defining covariates?

3.1 Level plot

Level plots are typically used to show geographic surfaces in a plane. In the subgroup analysis setting, two categorical variables are arranged in the axes, and the main plot area consist of cells that represent disjoint subgroups. Each subgroup is defined by the corresponding combination of levels of both covariates and a colour scale is used to display the treatment effect in that subgroup. In Figure 1 we show the implementation of a level plot for treatment effect in terms of log-hazard ratios in subgroups defined by age and weight for the prostate cancer dataset. Each covariate is partitioned into three levels. We additionally include the subgroups' sample sizes inside the cells. The cells on the bottom and the left margins represent the marginal subgroups corresponding to each of the three levels of age and weight, respectively. A divergent colour scale with range from -3 to 3 is used for the log hazard ratio. We also added the point estimate and confidence interval for the overall population in the legend.

This graphical approach is attractive since permits a direct and easy interpretation of effect sizes, therefore satisfying criterion C1. A quick look at the colours allows drawing conclusions such as for which subgroups the treatment is beneficial and for which ones is harmful. However, neither the overall treatment effect size nor the variability of the estimations can be represented in this plot, therefore making in it impractical to detect treatment effect heterogeneity. Although the addition of the sample sizes in the cells allows a comparison of the subgroup sizes, the sample sizes are not represented by the figure, therefore this display meets criterion C2 only partially. Level plots may only display pairwise overlay of marginal subgroups rather than all overlap across subgroups. It is worth noting that only two covariates can be considered in a level plot and although the number of the marginal subgroups of each covariate can be easily ten (therefore, the number of subgroups can reach to a hundred), this may lead to small subgroup sample sizes or even empty subgroups. Finally, because the cut-off points for continuous covariate are arbitrary, level plots are more suitable for categorical covariates.

Examining Figure 1, we may conclude that the treatment is actually worse for older patients and young patients with low weight. Moreover, the treatment seems to be even more beneficial for heavier young patients. However, these interpretations need to be taken with care, as the precision of the estimates is not given and the small sample sizes in the subgroups may lead to highly variable estimates.

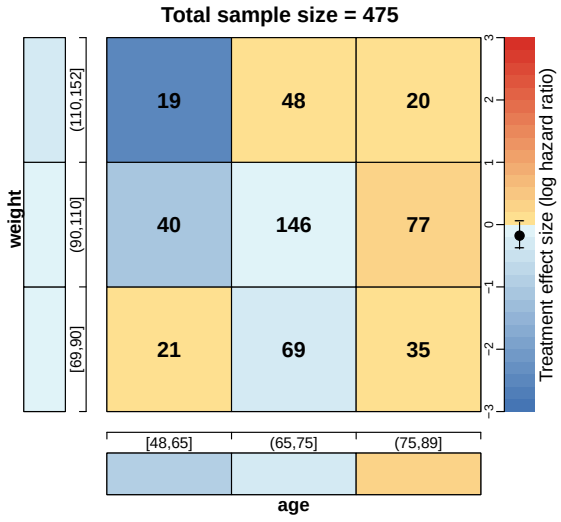


Figure 1: Level plot of treatment effect in terms of the hazard ratio across mutually disjoint subgroups defined by age and weight. The cells on the bottom and the left margins are the marginal subgroups corresponding to the levels of age and weight. The inner figures of the cells stand for the subgroup sample sizes.

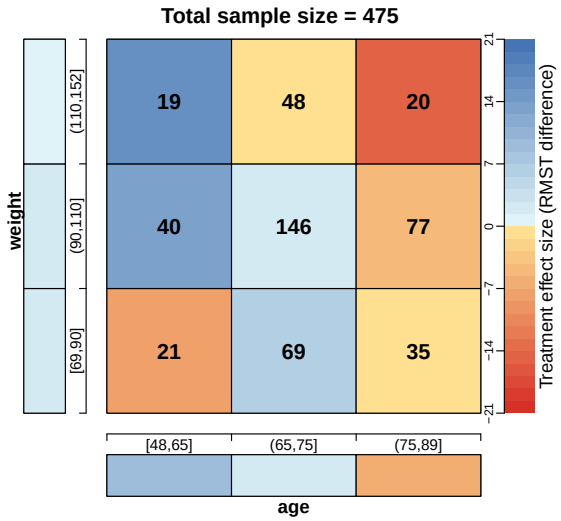


Figure 2: Level plot of treatment effect differences in terms of the restricted mean survival time (RMST) across mutually disjoint subgroups defined by age and weight. The cells on the bottom and the left margins are the marginal subgroups corresponding to the levels of age and weight. The inner figures of the cells stand for the subgroup sample sizes.

NB: Should we keep both the plot with HR and RMST? I would say only the one with HR.

3.1.1 Improved level plot

As a possible improvement, we propose to draw the area size of the coloured square inside each cell representing the proportion of the subgroup sample sizes relative to the full population (Figure 1). This new design feature allows one to compare subgroup sample sizes more easily. At the same time, it may be difficult to see the colour in each square, particularly in the case of small sample sizes. Perhaps a better way to present the information of the level plot is using a mosaic plot as described in the following section.

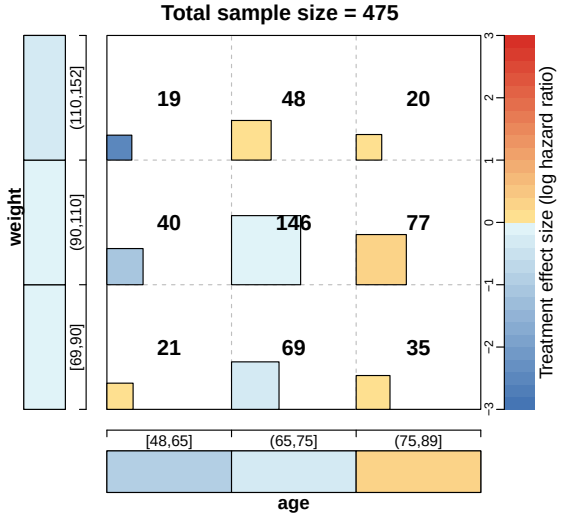


Figure 3: Level plot of treatment effect in terms of the hazard ratio across all mutually disjoint subgroups defined by age and weight, with sides of cells proportional to sample sizes. The cells on the bottom and the left margins are the marginal subgroups corresponding to the levels of age and weight. The area of the squares inside the cells are proportional to the sample sizes, which are also displayed in the middle of the cells.

3.2 Mosaic Plot

Mosaic plots are useful to represent contingency tables through arranging proportional-to-size cells in a grid. There are a number of variations in which this type of plot may be used in subgroup analysis. First, we devise an improvement of the level plot as in Figure 4. The interpretation of this plot is similar to the level plot presented in Figure 3.

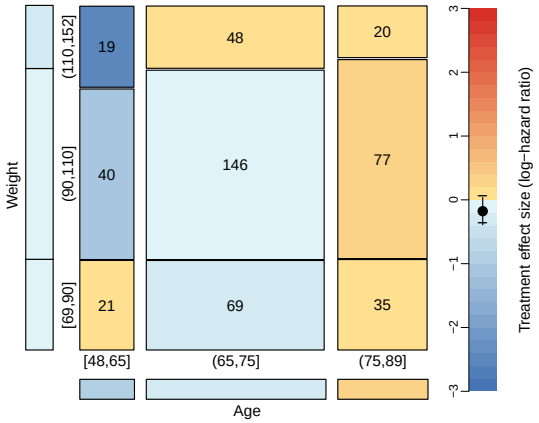


Figure 4: Mosaic plot displaying the subgroups conformed by the intersection of all subgroup defining covariates and their sizes.

Mosaic plots offer the advantage that a larger number of covariates can be arranged. In Figure 5, we used weight, performance and bone metastasis to illustrate a mosaic plot with three subgroup defining covariates. Note that a black line is drawn to show that there are no subjects in the subgroup defined by the higher weights, performance, and existence of bone metastasis. We interpret that there may be heterogeneity in the treatment effect when comparing subjects with and without bone metastasis when their weight is between 90 and 110 kg. and their performance is yes.

Lastly, we could use mosaic plots to illustrate event rates per treatment group across the levels of one subgroup defining covariate, as it is used in previous literature [7]. Although this plot may be more appropriate when the endpoint is binary, it is possible to adapt it for survival endpoint by using, for example, 2-year survival (Figure 6). In this case, it is possible to observe how survival rate is larger for treatment in the younger patients while the survival rate is larger for control in the older patients

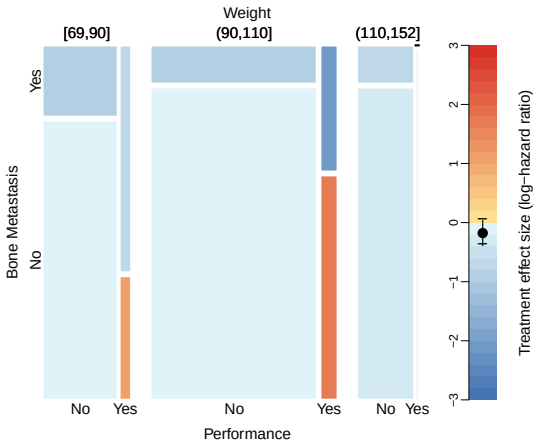


Figure 5: Mosaic plot displaying the subgroups conformed by the intersection of all subgroup defining covariates and their sizes.

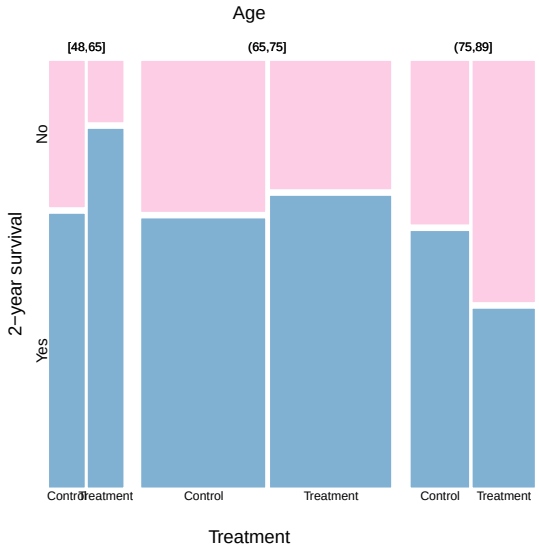


Figure 6: Mosaic plot displaying the subgroups conformed by the intersection of all subgroup defining covariates and their sizes.

3.3 Contour plot

The layout of a contour plot is similar to that of level plots: two variables are arranged along the axes. However, and in contrast to level plots, contour plots may allow to use continuous variables without categorizing them a priori. In this implementation, we define the subgroups by ranges of each covariate, specifying the desired sample sizes and possible overlaps. Each subgroup is then formed by neighbouring subjects in therms of their values on age and weight. For each subgroup, we calculate the hazard ratio for treatment vs. control, and place the estimate with its corresponding colour scale in the position of the subgroup's mean for each covariate. We draw contour lines through a bivariate interpolation and smooth surface fitting for irregularly distributed data points at pre-specified grid points. Figure 7 shows a contour plot for the treatment effect differences across age and weight. The points inform the corresponding subgroups effect sizes of the evaluated subgroups. We also use a divergent colour scale for the effect sizes in subgroups. Subgroup sample sizes and the overlap are adopted by design and therefore only annotated on top of the figure.

According to our assessment, contour plots match criteria C1 and C5 but not C2, C3 and C4. The total number of subgroups (corresponding to the number of points) can be more than ten by controlling the overlap proportions with neighbouring subgroups. However, there is no graphical display about subgroup sample sizes and overlap proportions. Such information can be only annotated in the subtitle and the caption the figure. Contour plots do not provide enough information to asses treatment effect heterogeneity because the overall treatment effect size and the precision of the estimates is not given.

There are few more noticeable characteristics for this graphical technique. Contour plots are particularly useful when a dataset size is rather large and for variables well distributed over the pre-specified rectangular region. This graphical approach only considers two continuous covariates. Moreover, the interpolated effect sizes may be unreliable in the region where only sparse points are irregularly distributed or no data point lies. In situations where the values of two covariates are sparsely distributed over the region, it may be unclear how smooth the interpolated surface should be. We also acknowledge that there may be other implementations of this plot. For example, it may be also possible to use local regression techniques to calculate the treatment effect at each coordinate.

In Figure 9 we display the same contour plot using filled areas, which resembles a level plot. The colour of each area correspond to the treatment effect size. This modification may make the interpretation easier. We observe a similar pattern as the one found in Figure 1, in which the older patients seem not to benefit from the new treatment. Again, this interpretation should be cautious as the precision of the estimates is not displayed.

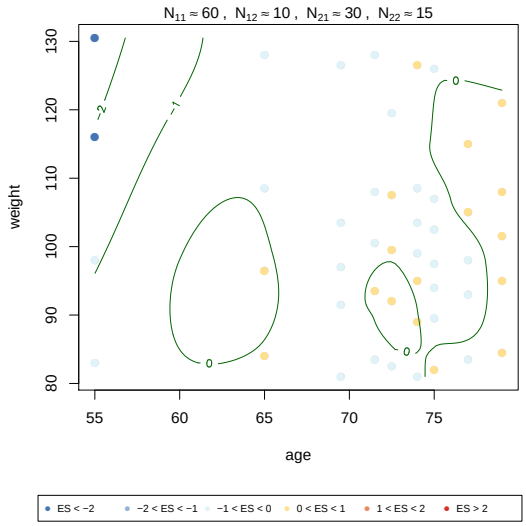


Figure 7: Contour plot of treatment effect in terms of the log-hazard ratio over the plane of age and weight. N_{11} stands for the sample size of a marginal subgroup defined by a range of age, N_{12} is the overlap size of the immediate marginal subgroups on age, N_{21} is the sample size of the subset of a marginal subgroup on age but further defined by a range of weight, and N_{22} is the overlap size of the immediate subgroups (which are the subset of a marginal subgroup on age) on weight.

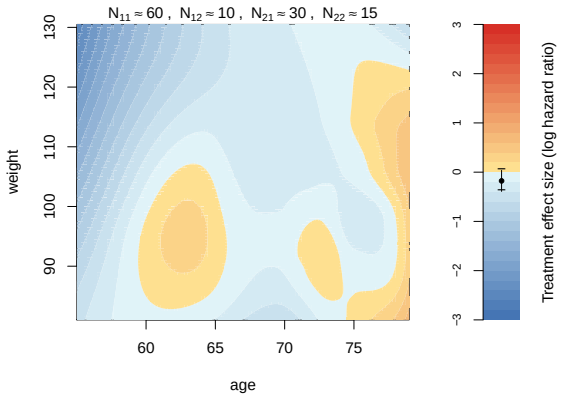


Figure 8: Filled contour plot of treatment effect in terms of the log-hazard ratio over the plane of age and weight. N_{11} stands for the sample size of a marginal subgroup defined by a range of age, N_{12} is the overlap size of the immediate marginal subgroups on age, N_{21} is the sample size of the subset of a marginal subgroup on age but further defined by a range of weight, and N_{22} is the overlap size of the immediate subgroups (which are the subset of a marginal subgroup on age) on weight.

3.4 Venn diagram

Venn diagrams are undoubtedly the most widely used tool to visualize sets and their relations. In the subgroup analysis setting, Venn diagrams are useful to display the composition of a dataset. A Venn diagram for subgroups defined by bone metastasis, history of cardiovascular events and performance is shown in Figure 9. Each circle defines the subgroup of patients for which the level of the corresponding variable is "yes". The diagram indicates the sample sizes for all the subsets that are formed by set operations (intersection and complement) on the three subgroup defining covariates. The number outside of the three circles indicates the size of the complement of the union of the three subgroups.

Venn diagrams satisfy C2, C3 in our assessment. Useful extensions to Venn diagrams, such as the Edwards' construction [8, 9], are available so that they can accommodate a large number of

subgroups, therefore also meeting criteria C5. The total number of subgroups including mutual disjoint ones can be 2^p, where p is the number of the sets considered. Despite this merit, there is a limit on the number of the sets considered in practice. It may become complicated to interpret a Venn diagram with more than five subgroup defining covariates.

Information about treatment effect differences in subgroups is unavailable.

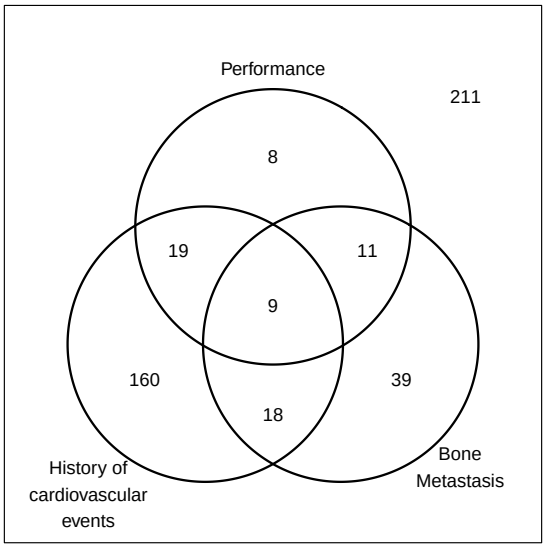


Figure 9: Venn diagram of 3 subgroups defined by presence of bone metastasis, history of cardiovascular events, and performance status = 1.

3.4.1 Improved Venn diagrams

NB: Do we need the unfilled venn diagram? maybe we can just start with the filled one.

Figure 10 and 11 are the improved Venn diagrams considering four and three subgroups, respectively. Both represent the treatment effect in terms of the log-hazard ratio of subgroups by colouring the corresponding regions, where the magnitude is shown in the colour bar on the right. This feature thus enables Venn diagram to satisfy the criterion C1. But it does not serve for detecting heterogeneity in subgroup effect sizes because the overall effect size is not given.

As seen in Figure 10, using four ellipses for representing all possible subgroups (formed through intersection and complement) is visually appropriate. Other patterns (such as polygons [10, 11]) can be also applied but the visualisations may not be easy to understand. In our example, however, we obtain very small subgroups when considering the intersections of the four covariates. The white regions indicate that it is not possible to calculate the treatment effect in the corresponding subgroup. There are two regions in which the log-hazard ratio takes very extreme values as the sample sizes of the subgroups are 4 and 5, which makes the estimate very unreliable. An additional rule may be added to this plot to colour only the areas that attain a pre-specified sample size.

Figure 11 further considers area-proportional methods, where each covariate representative region area is proportional to the respective sample size proportion. The region areas only approximately correspond the sample size proportions because of the limited degrees of freedom for circles. We employed the simple algorithm mentioned in[12]. In fact, other algorithms to display each region area proportional the sample sizes are available. Recently Micallef and Rodgers developed an algorithm that can produce an accurate area-proportional Venn diagrams using ellipses [12]. However, their algorithm is somewhat sophisticated and can only work on three sets.

Figure 11 shows that the treatment effect is reversed, being control better that treatment, for those subjects without bone metastasis when they have a history of cardiovascular events or their performance is yes.

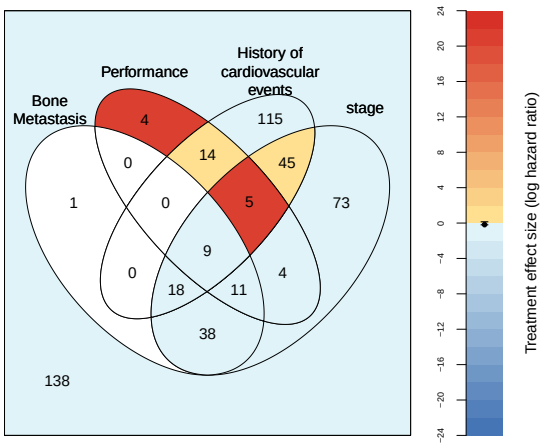


Figure 10: Venn diagram of 4 sets defined by presence of bone metastasis (bm), disease stage, performance status = 1 (pf) and history of cardiovascular events (hx) with treatment effect sizes in terms of the log-hazard ratios.

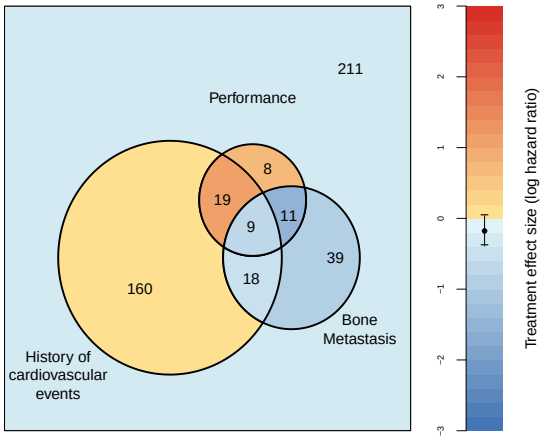


Figure 11: Approximate area-proportional Venn diagram of 3 subgroups defined by presence bone metastasis (bm), history of cardiovascular events (hx) and performance status = 1 (pf) with treatment effect sizes in terms of the log-hazard ratios.

3.5 Bar chart

Another useful graphical technique to depict treatment effect sizes are bar charts. Bar charts are easy to interpret and allow a direct comparison among subgroups. For the subgroup analysis problem, we use subgroups defined by the levels of two categorical variables, although other constructions may also be available. A bar chart for treatment effect differences in subgroups defined by age and weight is shown in Figure 12. Each covariate is categorized into three levels and the bars represent mutually disjoint subgroups. The levels of age and weight are respectively listed in the top and the bottom part of the picture. The height of the bars are proportional to the treatment effect differences between the treatment/control arms, that is, the difference in RMST. The width of the bars are proportional to the square root of the subgroup sample sizes. This arrangement therefore has also another useful property: the area of the bars is proportional to the restricted mean survival gain or loss when using treatment in comparison to control. The colours of the bars merely shows which subgroup has the same category level on age.

Based on our assessment, this graphical representation approach holds C1, C2 and C5, partially C3 but not C4. Each bar is the pairwise overlap of two subgroups defined by age and weight with their respective levels. Therefore, bar charts only provide partial overlay information. Such a graphical approach does not allow to examine heterogeneity in treatment effect differences across subgroups due to no display of the overall effect size. In terms of C5, bar charts can handle more than ten (mutually disjoint) subgroups if the number of levels of each covariates are larger.

Few noteworthy characteristics also need to be mentioned. First, bar charts exhibit the standard errors of the point estimator for subgroup effect sizes. Second, it only considers two subgroup defining covariates. If considering few more covariates, one could label all the level combinations of the covariates in the bottom part of the picture or simply to make a legend elsewhere. Third, although it satisfies C5, a high number of covariates or levels may be problematic, making it difficult to compare the widths of the bars. Fourth, as in level plots, the cut-off points for categories in continuous variables may be arbitrary and bar plots is therefore preferable for categorical covariates.

Although we use a different measure for the treatment effect, the direction of the estimates is maintained compared to the level plot in Figure 1 and the interpretation remains unchanged.

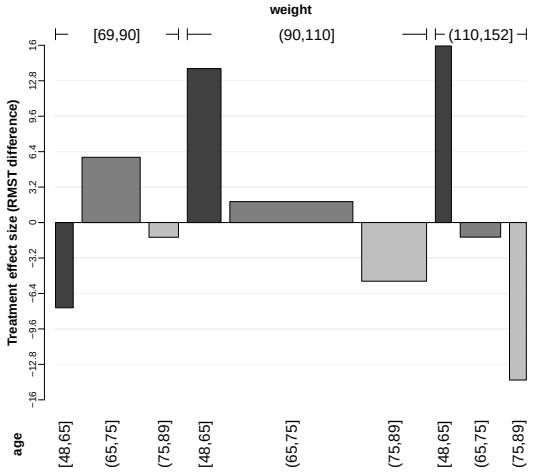


Figure 12: Bar chart of 9 mutually disjoint subgroups defined by the levels of age and weight

3.6 Forest plot

Although forest plots are a common graphical display approach for meta analysis, they are also extensively used for subgroup analysis. In a forest plot, the treatment effect estimates along with their confidence intervals for the subgroups defined by a number of covariates are displayed vertically. The overall treatment effect is also plotted on top allowing a direct comparison. It is also suggested that a vertical line at the overall treatment effect level is added to facilitate seeing if a subgroup confidence interval differs significantly from the overall effect [13]. Additional information in a table format is usually included to provide the exact magnitude of the estimates. Figure 13 shows its applications to the estimation of treatment effect differences in eight subgroups defined by four binary covariates. The text on the left panel shows the mean estimate of treatment effect difference, lower/upper bounds of 95% C.I and subgroup sample sizes (further divided into treatment group and control arms). The middle panel displays the treatment effects with their confidence intervals. The squares in the middle are proportional to the subgroup sample sizes When using a continuous or binary endpoint, it is also recommended to include the effect estimates for treatment and control to observe whether both interventions have harmful effects despite the promising effect size. In our implementation for survival endpoint, we include the Kaplan-Meier curves for each subgroup.

From the above description, forest plots in the subgroup analysis setting hold all the criteria but C3 because of the inability to show subgroup overlaps. Note that the visual judgement on heterogeneity is slightly different from those in [1, 13, 14]. We later adopt the same recognition rule for the graphical approaches with similar design features.

We observe in Figure 13 that the subgroup with bone metastasis is the subgroup with the largest benefit from the treatment, while for the rest of the subgroups their treatment effect is closer to that on the overall population. The Kaplan-Meier curves allow to rapidly recognize the differential pattern in the subgroup with bone metastasis.

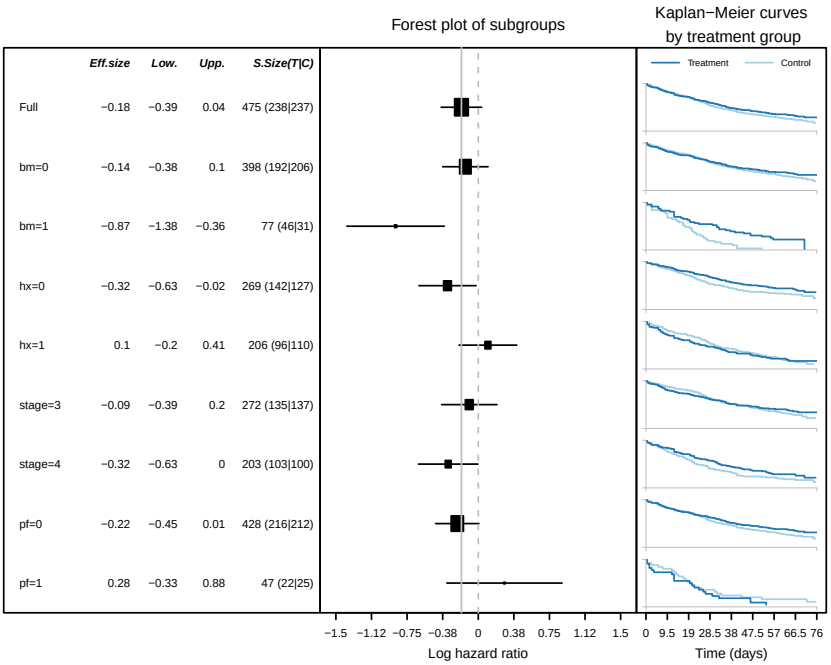


Figure 13: Forest plot for subgroups defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes in terms of the log-hazard ratio and associated treatment and control group Kaplan-Meier curves are displayed.

3.7 Tree plot

The tree plot for subgroup analysis starts with the full population that branches into two or more items, corresponding to the levels of the first subgroup defining covariate. Each of the items in the new level branch again into two or more levels for the second covariates, then for the third and so on. If more variables were included, this division procedure is consecutively conducted to form subgroups until all the category combinations of the covariates are considered. Figure 14 shows a tree plot of treatment effect differences for subgroups defined by bm, pf and hx. In each layer, treatment effect differences and their 95% confidence intervals (C.I.) for the associated subgroups are also displayed. The purple horizontal lines placed in the middle of C.I. have a length proportional to the weight of subgroup sample size over the full population. An additional horizontal dotted line is added for the overall treatment effect size.

Tree plots match all the criteria. It is obviously fit to C1, C2 by design. In addition, the subgroups at the each layer are formed by the intersection of the levels of the covariate in that layer with the covariates that are placed above them, thus holds C3 for displaying the information of all subgroup overlaps. Moreover, examining heterogeneity in treatment effect differences of subgroup can be fulfilled for C4. Similar to forest plots, the assessment demands drawing an auxiliary horizontal line with the y-coordinate at the overall effect size for each layer and then seeing whether there is any C.I. not crossing the line. As to C5, tree plots can certainly address more than 10 subgroups. Note that the number of subgroups also depends on how many covariates are involved and how many categories each covariate has.

A few features of tree plots are worthily pointed out. First, it provides information of the interval estimation for subgroup effect sizes. Second, it is possible to consider more than two categories for each covariate if needed. Ideally, however, the number of covariates and categories should be moderate of we may end up with subgroups small sample sizes. Finally, when considering continuous covariates tree plots have the same issue about arbitrary cut-off points as level plots and bar charts

Figure 14 allows us to draw additional conclusions regarding the treatment effect sizes. We continue to observe that the treatment effect is more pronounced for subjects with bone metastasis. Additionally, we notice that the subgroup without bone metastasis but with history of cardiovascular event and performance status 1 has a positive log-hazard ratio, implying that the control is better than treatment for this subgroup.

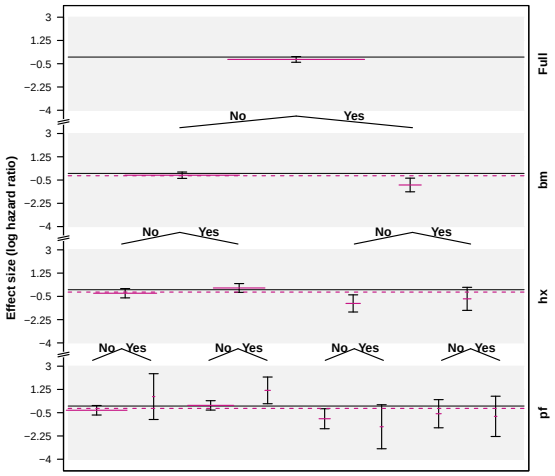


Figure 14: Tree plot of treatment effect difference for subgroups defined by all category combinations of the covariates existence of bone metastasis (bm), history of cardiovascular events (hx), and performance (pf). Each layer shows the 95% C.I. of treatment effect differences for the associated subgroups. The purple horizontal lines placed in the middle of C.I. have a length proportional to the weight of subgroup sample size over the full population.

3.8 Galbraith plot

A Galbraith plot [15, 16] is an alternative or supplementary to a forest plot for examining heterogeneity of studies or subgroups in meta analysis. Its variant shown in Figure 15 exhibits the estimation of treatment effect sizes for subgroups defined by the four binary covariates covariates. The xy-coordinates correspond to the points:

$$x_i = 1/\sqrt{\mathrm{Var}(\hat{\delta}_i - \hat{\delta}_F)}, \quad y_i = (\hat{\delta}_i - \hat{\delta}_F)/\sqrt{\mathrm{Var}(\hat{\delta}_i - \hat{\delta}_F)} \tag{1}$$

where $\hat{\delta}_F$ is the treatment effect estimate in the full population and $\hat{\delta}_i$ is the treatment effect estimate in subgroup i, $i = 1, ..., K$. The grey band serves to examine heterogeneity if one standardized estimate is located outside the band and the arc indicates effect sizes. The effect size estimate of a subgroup is registered at the red icon projected on the arc by the line from the origin through the corresponding point. The central line points to the average effect effect for the full population. See Appendix A for details on the calculation of the points.

The result of the graphical assessment of Galbraith plots is satisfactory. It obviously holds C1, C4 and C5 because of its design features. It can handle a large number subgroups and is also helpful to detect outliers. Galbraith plots only partially fit to the criterion C2, since it only indirectly reveals information of subgroup sample sizes through individual standard errors. Moreover, it does not hold C3. Like forest plots, it is not possible to know subgroup overlap information.

In terms of our example, we conclude that treatment effect heterogeneity may be present in the subgroup of patients with bone metastasis and in the subgroup of patients with history of cardiovascular effects.

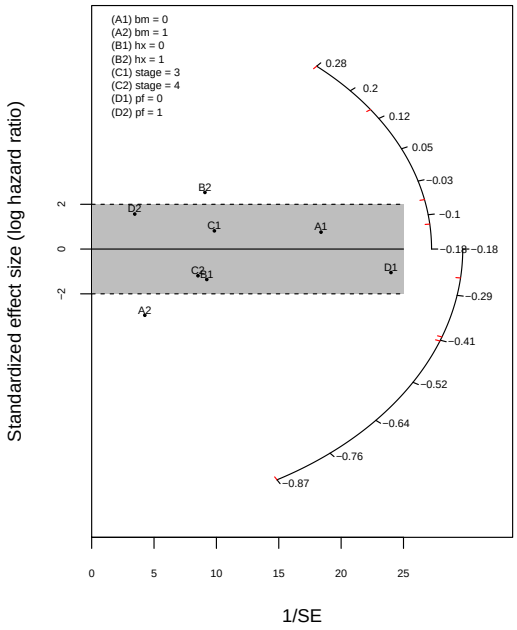


Figure 15: Galbraith plot across subgroups and their complements defined by stage, history of cardiovascular events (hx) and existence of bone metastasis (bm)

3.9 L'Abbé plot

L'Abbé plots [17] are a variant of scatter plots which is useful for examining heterogeneity in meta analysis. The graphical design is originally for binary outcome data to represent risk ratios, risk differences or odds ratios between treatment and control arms. For our implementation, we extend this graphical technique to the case of continuous and survival outcomes and also modify points to rectangles (Figure 16). Each subgroup has one estimate located at a position where its x-coordinate and y-coordinate correspond to the estimates of the RMST in control and treatment arms, respectively. The width and the height of a rectangle (corresponding to a subgroup) respectively indicate the sample sizes of control group and treatment group. We draw a diagonal dashed line that represents no treatment effect (equal RMST in both arms) and a solid diagonal line with y-intercept at the overall treatment effect size. Each rectangle has a vertical segment from its centre to the diagonal dash line, representing the gain (if blue) or loss (if red) in terms of RMST when comparing treatment vs. control The subgroup treatment effect sizes are written in the top-left corner of the picture. The other vertical purple dashed lines which start on the diagonal line have the lengths same as the upper or lower bound of 95% C.I. of the effect sizes for subgroups. If one vertical purple dashed line does not cross the solid line, heterogeneity in subgroup effect sizes may occur. This design feature is similar to that in forest plots. NB: Should we keep this line? I think it is confusing and too much for the plot as it overlaps with the other coloured line. See alternative in Figure 17

L'Abbé plots share the same graphical assessment results with forest plots. They satisfy all criteria except C3 for not showing subgroup overlap information. Two characteristics should be noted. First, while they may handle many subgroups, it may be difficult to recognize the corresponding rectangles if subgroups have a close effect estimate for treatment and control groups. Second, it does not fully reveal information about interval estimation of subgroup effect sizes and of treatment effects in treatment/control subgroups

This graphical tool allows us to draw additional conclusion in our example. The subjects with bone metastasis in the control group have a lower RMST. When receiving control, however, the RMST is closer to that of the other subgroups.

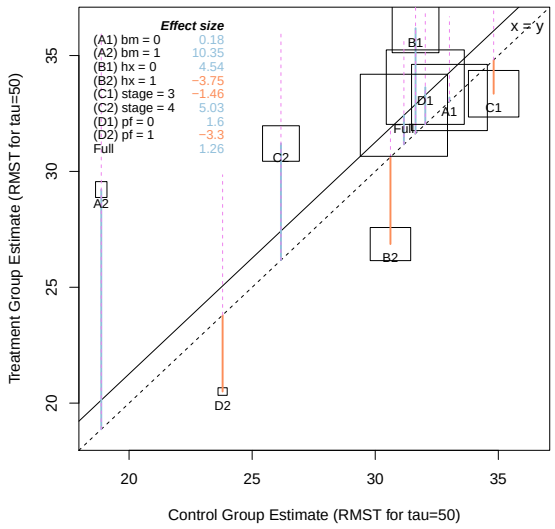


Figure 16: L'Abbé plot for the subgroups and their complements defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes are given in terms of the difference in restricted mean survival time (RMST).

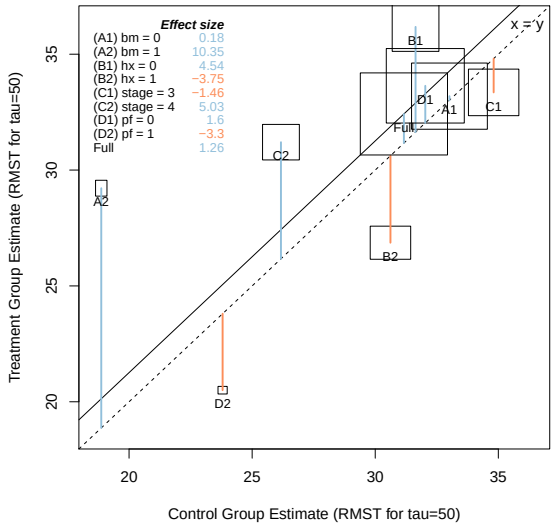


Figure 17: L'Abbé plot for the subgroups and their complements defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes are given in terms of the difference in restricted mean survival time (RMST).

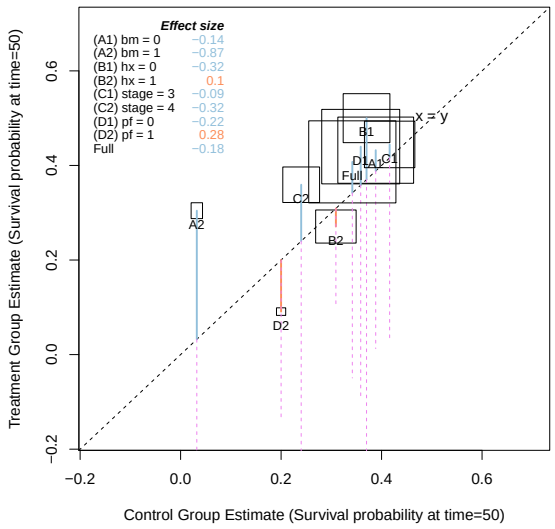


Figure 18: L'Abbé plot for the subgroups and their complements defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes are given in terms of the log-hazard ratio.

3.10 STEPP

The sub-population treatment effect pattern plot (STEPP) [18, 19] is of some publicity in breast cancer recently. It is a non-parametric method mainly for examining whether treatment-covariate interactions exist. In Figure 19, we adopted the slide-window fashion of STEPP to represent the estimation of treatment effect size (log-hazard ratio) in overlapping subgroups defined by age. Each subgroup has a sample size of around 40 and also has about 80% being overlapped with the neighbouring subgroups. The band bounded by the blue dashed lines is constructed for 95% simultaneous confidence interval (C.I.). The other band bounded by the orange dashed lines is built based on individual 95% C.I. (without multiplicity adjustment). The red line is formed by connecting the mean point estimates of treatment effect difference for all individual subgroups. The green line represents the log-hazard ratio estimate for the full patient population. It is noted that the point estimates (including mean, the boundaries by 95% simultaneous C.I. and individual C.I.) are marked in the middle of the interval defined as a subgroup. If the green line does not lie in the region formed by simultaneous confidence intervals, it reveals that interaction may exists.

The STEPP approach has a reasonable graphical assessment result, as it matches C1, C4 and C5. Here the information about subgroup overlap and sample sizes is only annotated in the figure and the caption. It is noted that the number of subgroups depends on the sample size of subgroups and the overlap proportions.

This plot only considers one continuous covariate. It is difficult to extend the application for more continuous covariates. The subgroup sample sizes should be specified by design, and in some situations a researcher may have no clear idea about how large a subgroup should be and how much it should overlap with the immediate subgroups. Perhaps, practitioners need to conduct sensitivity analysis for a different sample sizes for subgroups and overlaps. The analysis results may further be compared with the graphical results by using MFPI algorithm [20, 21] or non-parametric methods (such as Gaussion processes [22]), where a functional curve of the covariate on treatment effect is interpolated.

For our example, we observe that the treatment effect for subgroups defined by age fluctuate around the overall treatment effect. When approaching the ends of the range of the covariate the estimate of the log-hazard ratio departs form the estimate for the full population, although the confidence intervals still cover it.

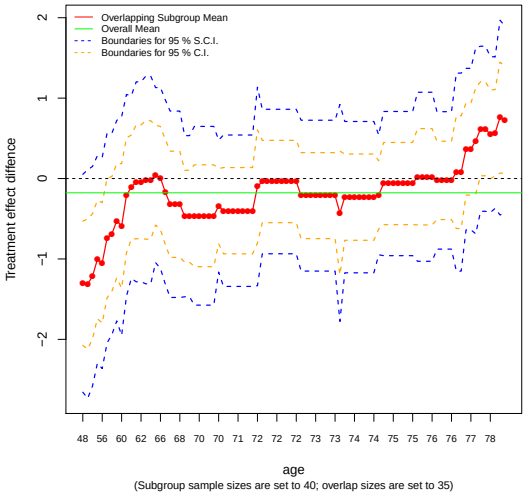


Figure 19: STEPP plot of overlapping subgroups defined by age. Each subgroup has a sample size of around 40 (N11 = 40) and is controlled to have about 87% (N12/N11) being overlapped with the neighboring subgroups.

3.11 Alluvial diagram

Alluvial diagrams are flow diagrams that can be used to display the distribution of the subjects across the subgroup defining covariates. Figure 20 shows one possible implementation of the alluvial plot for the subjects in the prostate cancer dataset using the `alluvial` R package. The blue coloured bands correspond to patients that were randomized to treatment while light-blue bands to patients in control. The height of the bars for each category in the subgroup defining covariates is proportional to the numbers of subjects in this category, therefore giving a notion of the size of the subgroup. Each alluvia (or band) represents the combination of values for the covariates. Therefore this diagram has also the advantage of giving an idea of the overlap of the subgroups, via the width of the alluvium (or bands).

As the mosaic plots, alluvial diagrams may also be used to illustrate event rates per treatment group across the levels of the subgroup defining covariate. Figure 21 shows the 2-year survival per treatment arm across levels of performance, history of cardiovascular events and bone metastasis. We also modify this plot to display it vertically, which may ease the interpretation of the diagram.

Alluvial diagrams do not provide any information regarding treatment effect sizes, but only on the composition of the subgroups, meeting criteria C2 and C3 as Venn diagrams. Alluvial diagrams can also display a large number of subgroups and can be used not only with binary covariates but also categorical ones. When the covariates are continuous however, parallel coordinates plots can be used in a similar way.

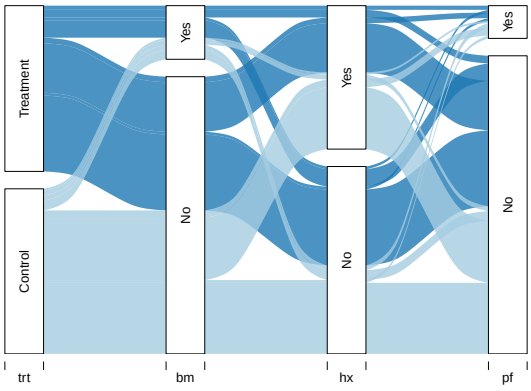


Figure 20: Alluvial diagram displaying the distribution of patients across the subgroups. The red coloured bands correspond to patients that were randomized to treatment while blue bands to patients in control. The width of the bands are proportional to the sizes of the subgroups.

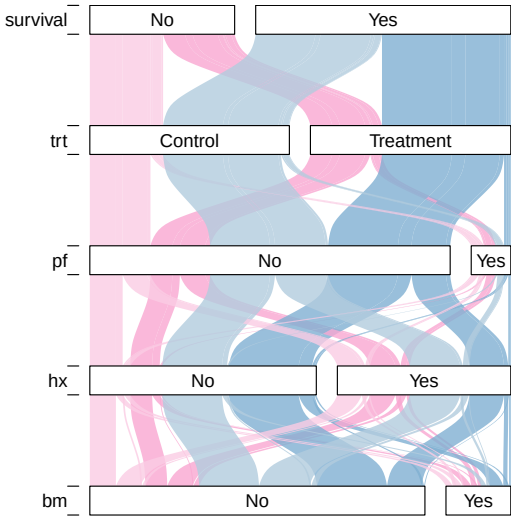


Figure 21: Alluvial diagram displaying the distribution of patients across the subgroups. The red coloured bands correspond to patients that were randomized to treatment while blue bands to patients in control. The width of the bands are proportional to the sizes of the subgroups.

3.12 UpSet Plot

UpSet plots are a novel visualization technique for the quantitative analysis of sets and their intersections [23]. It was proposed to overcome the limitation of Venn diagrams of showing up to a small number of sets or subgroup defining covariates. In Figure 22, we used the `UpSetR` package [24] to create the plot with use the six subgroup defining covariates (age is dichotomized >75 years, and weight >100). The sizes of the univariate subgroups for these covariates are shown in the horizontal bar plot at the bottom-left corner of the figure. The "matrix" layout allows visualizing the intersection of the covariates and main bar plot displays the sizes of the subgroups that are defined by these intersections. For example. the first bar indicates there are 58 subjects with age > 75, no history of cardiovascular events, disease stage 3, weight ≤ 100, no existence of bone metastases and performance status 0. Moreover, we added a 'query' to display the frequency of treatment and control in each subset.

Similarly to Venn diagrams, UpSet plots meet criteria C2 and C3. Additionally, UpSet plots also meet criteria C5, since their advantage is that they are scalable, and thus allowing a large number of subgroup defining covariates. It is not possible however, to display information on the effect sizes, although a modification to this plot is provided in Section 3.13. Another issue of this implementation is that the subgroups are defined by the intersections of all 6 covariates (the dots in the matrix panel indicates 0 or 1). In some cases, we may be interested in displaying the information of the intersection of a smaller number of covariates, marginally across the others. The UpSet web tool provides this option [25].

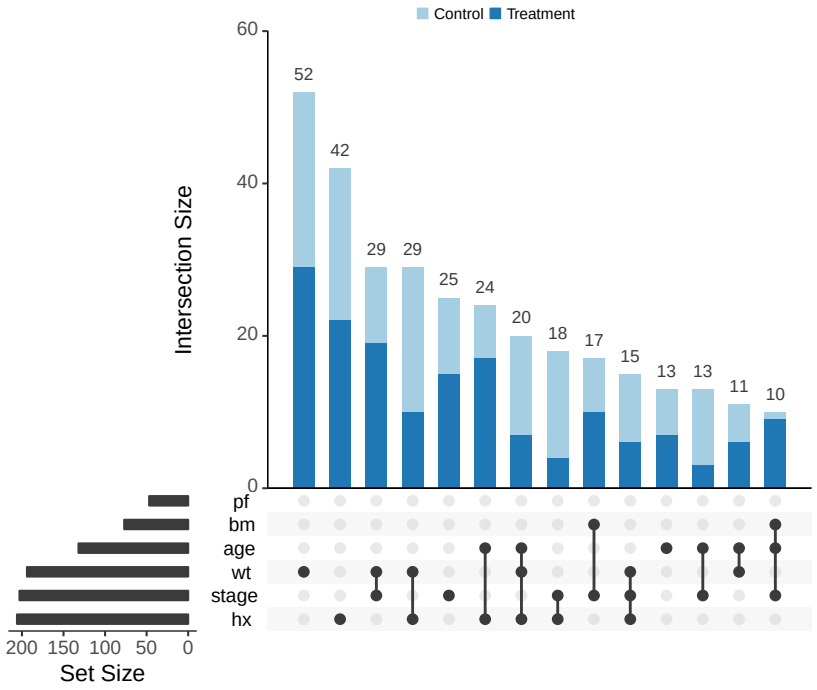


Figure 22: Upset plot displaying the subgroups conformed by the intersection of all subgroup defining covariates and their sizes.

3.13 Improved UpSet plot

We extend the UpSet package to display effect sizes in an extra panel (Figure 23). In this case, the log hazard ratio and its confidence interval is shown. This information is similar as that on the forest plots. However, the UpSet plot provides the advantage to observe intersection of sets and arrange them in terms of their sizes.

Our extension of the UpSet plot also allows to display lower level intersections. We implement a new icon for the matrix panel: a '+' symbol if variable is equal to 1 or 'yes', a '-' if variable is equal to 0 or 'no', and empty if this variable is not considered for the subgroup definition. For example, the first bar of the plot corresponds to the entire dataset, which has a size of 475. The second bar with a size of 428 corresponds to the subgroup of pf=0, irrespective of the values of the other two variables. Since the number of subgroups to consider increases dramatically in this modification (3^p subgroups when considering p binary covariates), only three covariates are considered. One can include, however, more covariates and filter the number subgroups according to different criteria, such as total sample size, sample size per treatment, etc.

As the overall treatment effect and its confidence interval is also included in this modification, it allows to compare treatment effects and check for treatment effect heterogeneity.

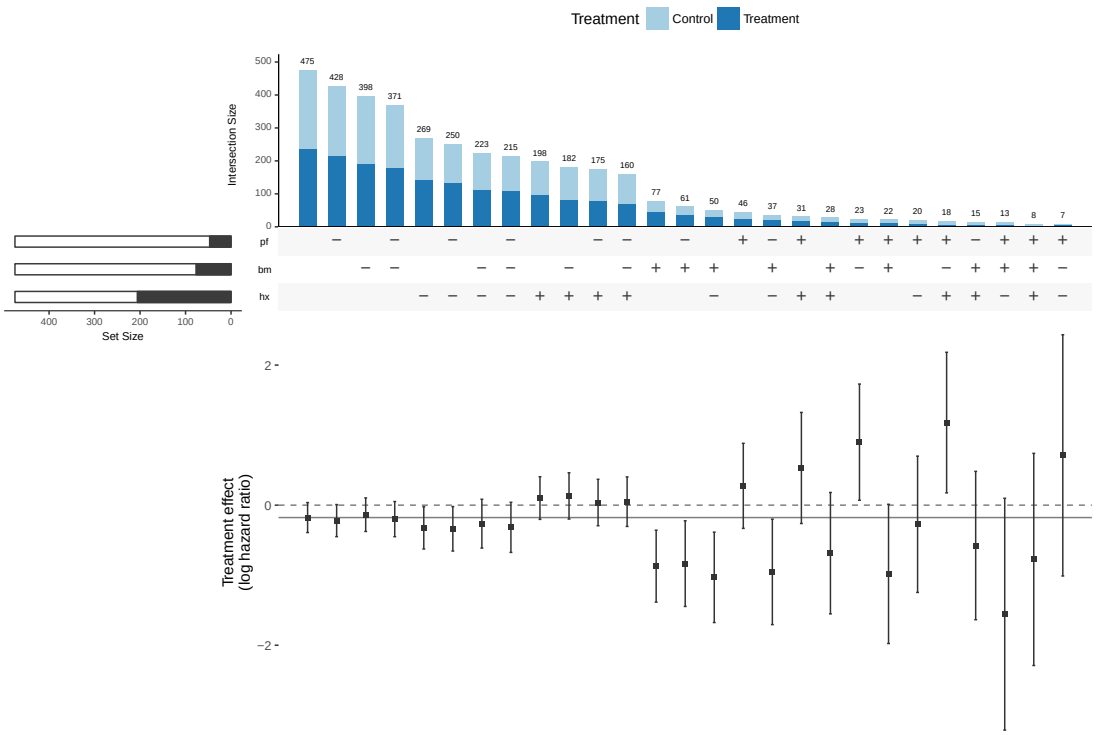


Figure 23: Improved UpSet plot for subgroups defined by performance (pf), bone metastasis (bm) and history of cardiovascular events (hx). Subgroups sizes are displayed on the panel on top with the bar plot, while the treatment effect sizes in the panel on the bottom. The panel in the middle displays how the subgroups are conformed by assigning a '+' if the variable is equal to 1 and a '−' if the variable is equal to 0

3.14 Circle Plot

Circular diagrams are widely used to visualize genomic data [26]. There are several approaches for the use of these diagrams, although the main aspect is that it allows to represent the relationships between pairs of sets. For our example, we use the categorized variables age and weight (Figure 24). The categories of each variable are arranged along the circle, where each of their corresponding cell have a size proportional to their sample size and a colour representing the treatment effect estimate, in terms of the log-hazard ratio. The ribbons on the centre of the diagram represent the overlap between the categories of the variables.

Circle plots meet all the criteria but C4 for displaying treatment effect heterogeneity. The flexibility of this plot is also an advantage, since many other implementations may be devised, specially when the number of covariates is extremely large as when dealing with genomic data.

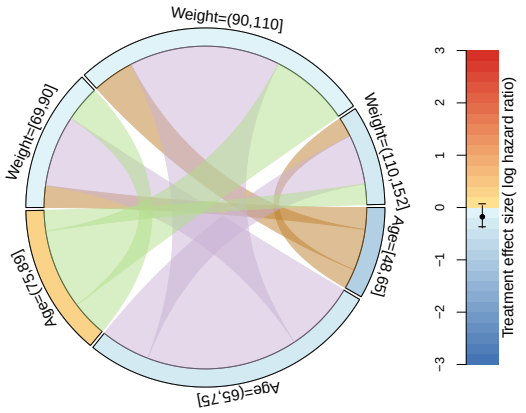


Figure 24: Circle plot displaying the subgroups conformed by the intersection of all subgroup defining covariates and their sizes.

4 Other alternatives

Forest plots, Galbraith plots and L'Abbé plots share the inability of showing subgroup overlaps. One potential improvement is to consider combining relevant figures about overlap information.

The plots shown in Figure 25 exhibit certain subgroup information about pairwise overlap proportions or similarity measures. Figures 25a- 25d show pairwise relative overlap proportions, where different colours show the range of overlap magnitude.

More specifically, Figure 25a is a plot with bidirectional arrowed curves. The position of arrows additionally indicates the information about how to calculate the relative overlap proportions. The subgroup labelled at the starting point is used as a baseline for calculating the relative proportion of the overlapping subgroup. Figure 25b is a variant of Figure 25a. Two identical sets of subgroup labels around two circles and each shows relative overlapping proportions with unidirectional arrowed coloured lines. The subgroup labelled at the starting point of the arrowed line is a baseline subgroup for the relative overlapping proportion. Figure 25c is a plot merely using coloured lines connecting subgroup labels on different levels. A subgroup label on the higher level is the baseline subgroup for the relative overlapping proportions with its counterpart on the lower level. Figure 25d is a matrix plot for relative overlapping proportions of pairwise subgroups. The row subgroup label indexes what subgroup should be as a baseline and the sizes of the circles signal overlap magnitude.

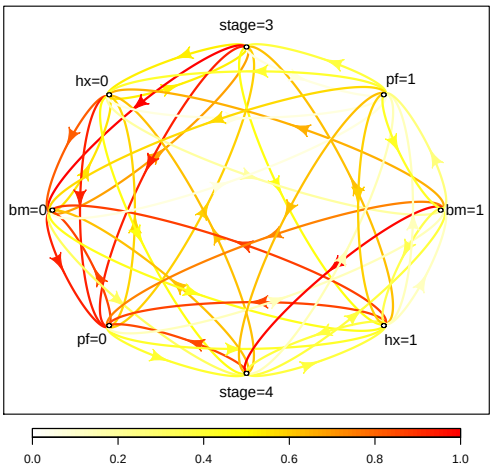
Both Figures 25e-25f show dissimilarity distance, which is defined by one minus a relative overlap proportion. Each line of Figure 25e shows the dissimilarity distance of a subgroup with the others. The red crosses along each line are located according to actual dissimilarity distances; the red subgroup labels correspond to the red crosses, where the labels are placed by order based on their actual dissimilarity distances. Figure 25f shows the same information as Figure 25e, where the coloured lines represent subgroups. There are one variant of Figure 25e shown in appendix. Note that for each subgroup we do not show its dissimilarity distance to itself and its complement.

Incidentally, the Jaccard index, namely $|A \cap B|/|A \cup B|$ for any sets A, B, can replace pairwise overlap proportions for subgroup overlap information. The graphical display is thus simplified due to not showing repetitive Jaccard indexes. However, this measure may lead to missing some information about whether a subgroup contains the others or not.

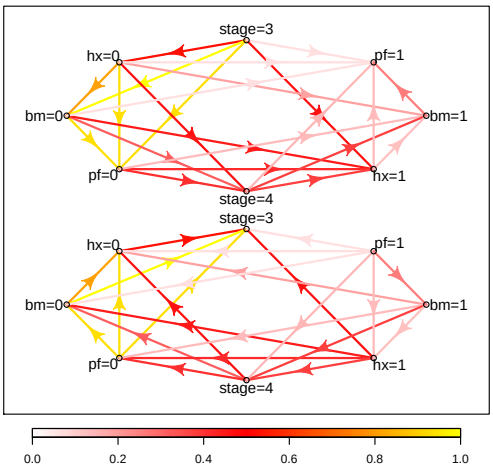
5 Discussions and conclusion

We have exploited several graphical approaches and assessed their characteristics for subgroup problems. We also attempted to improve some methods by mitigating their demerits. The assessment and characteristics of the improved approaches are summarised in Table 1.

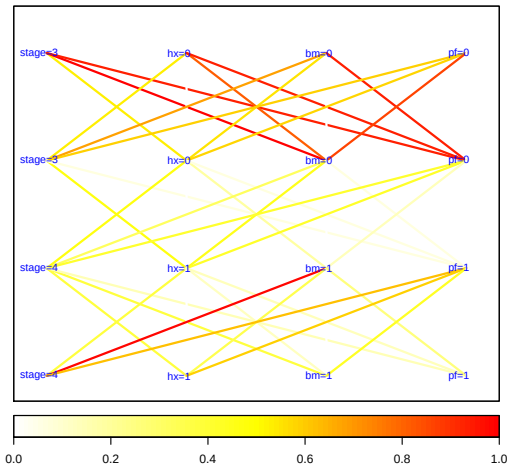
The general summary is as follows: most of the graphical techniques satisfy the primary criterion about displaying subgroup effect sizes. Except Level plot, contour plot and venn diagram, the rest displays or has information to construct confidence intervals, specifically bar chart and Galbraith plot exhibit standard errors for estimators. Furthermore, only forest plot and L'abbé



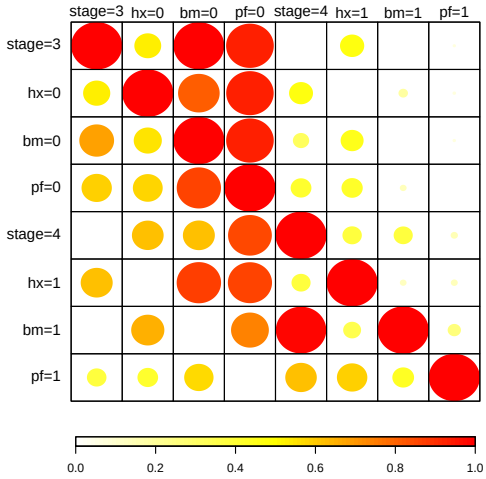
(a) Line plots with bidirectional arrowed curves for relative overlap proportions for pairwise subgroups.



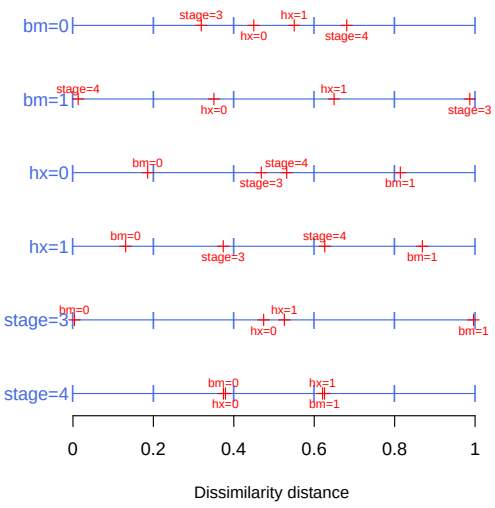
(b) Line plots with unidirectional arrowed lines for relative overlap proportions for pairwise subgroups.



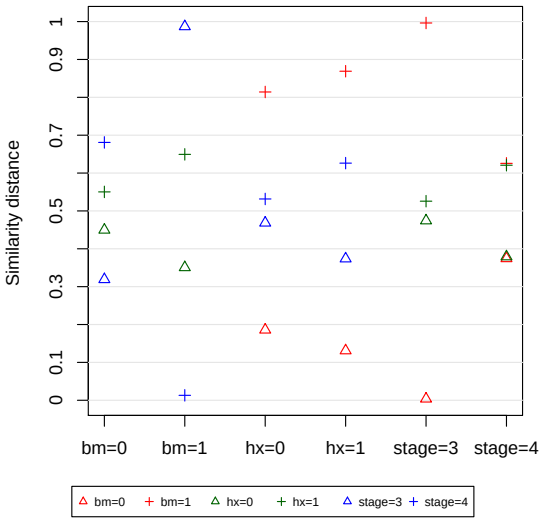
(c) Line plots for relative overlap proportions for pairwise subgroups.



(d) Matrix plots for relative overlap proportions for pairwise subgroups.



(e) Line plots 1 for dissimilarity measures.



(f) Line plots 2 for dissimilarity measures.

Figure 25: Plots for subgroup information about pairwise overlap proportions or dissimilarity measure.

Table 1: The assessment summary of graphical techniques for subgroup problems. The assessment criteria are: **C1**: whether to display effect sizes for subgroups; **C2**: whether to show subgroup sample sizes; **C3**: whether to exhibit all subgroup overlap information; **C4**: whether to serve for detecting heterogeneity in subgroup effect sizes (or the treatment-covariate interaction); **C5**: whether is available for the large number of subgroups (more than 10). The subscript * of some graphical approaches denote they have been improved. The C.I. column indicates whether the confidence intervals is provided in the plot or at elast the precision in the estimates. The overlap column corresponds to P: pairwise overlap or A: all overlap. N_c represents the number of covariates for considerations NB: there was a Δ symbol before in some of the plots, but not included in the legend to check what it stands for

	Criterion					Additional features			
	C1	C2	C3	C4	C5	C.I.	T/C Effect	Overlap	N_c
Level plot	✓	✓	✓		✓			P	2
Mosaic plot	✓	✓	✓		✓			P	2
Contour plot	✓		✓		✓			P	2
Venn diagram*	✓	✓	✓		✓			A	2-6
Bar chart	✓	✓	✓		✓	✓		P	1-5
Tree plot	✓	✓	✓	✓	✓	✓		A	1-5
Forest plot*	✓	✓		✓	✓	✓	✓	P	1-40
Gailbraith plot*	✓	✓		✓	✓	✓		P	1-100
L'Abbé plot*	✓	✓		✓	✓	✓	✓	P	1-40
STEPP	✓			✓	✓	✓		P	1
Alluvial		✓	✓		✓			A	1-10
UpSet*	✓	✓	✓	✓	✓	✓		A	1-100
Circle	✓	✓	✓		✓			A	2-100

plot further provide subgroup effect sizes for the treatment and control arms. In terms of the second criterion, the majority of the approaches provide a visual display on subgroup sample sizes. Only Galbraith plot indirectly show the information through the standard error of estimators.

The third criterion is fully and partially hold for all apart from contour plot and STEPP. Venn diagrams, tree plots, and UpSet show the overlay of all subgroups. The remaining approaches only display the overlay for pairwise subgroups. Six graphical displays featuring different design characteristics were invented for improving Forest plot, Galbraith plot, L'Abbeé plot and UpSet. It is noted that when the number of subgroup is small (say, up to five), the improved forest plot, Galbraith plot and L'Abbé plot can combine a Venn diagram for displaying subgroup overlap completely.

The capacity of detecting heterogeneity or interaction is equipped in the last five approaches. These five commonly feature a reference line corresponding to the overall effect size. Their judgement of heterogeneity generally depends on the distances between the line and subgroups or the location of the line within the confidence band. As for the last criterion, all the techniques can be available to handle more than ten subgroups. In particular, Venn diagrams and tree plots practically can deal with only up to five sets (considered for overlap) for effective visualisation. Even an area-proportional Venn diagram can afford merely three sets. Moreover, six approaches are able to regard a small number of covariates for subgroups. Only forest plots, Galbraith plots, L'Abbé plots, UpSet, and circle plots can deal with a middle or large number.

Although the assessment suggests the superiority of certain approaches, in practice, the decision of a technique for use still demands considerations of different characteristics and circumstances. For example, contour plots can be particularly useful when a data set is large and the distributions of two covariates considered are roughly uniform; level plots and bar charts may be easier for some audience to understand subgroup information due to their simple design; forest plots and L'Abbé plots can be used in the exploratory setting, especially to prevent the subgroup with adverse effects in both interventions despite the positive effect size; STEPP could be suitable for investigating the treatment-covariate interaction or exploring potential subgroups with positive findings if the covariate of interest is confirmed to impact treatment effect by other studies.

The approaches are worthy of further discussions in design and use issues. One is that the results of statistical inference based on hypothesis testing are not informed. Our primary goal is to visualise essential subgroup information including effect sizes and sample sizes. We consider all the approaches mainly serve as graphical descriptive tools, and therefore there is no need for adding the testing results for initial subgroup analyses. As a result, the presence of the positive and adverse findings in subgroups with small sample sizes only brings concerns to practitioners

for further investigations.

Another issue is correlation between categorical variables considered. The graphical approaches are not designed to address the problem that the correlation causes, where estimates from mutually disjoint subgroups can be correlated and thereby this may lead to confounding interpretations of subgroup effect sizes. This can be solved by using the standardization technique [27] in epidemiology before utilising the graphical approaches.

In addition, the focus on developing a two-dimensional graphical display can be contentious. We recognize the usefulness of other graphical alternatives including a three-dimension graphical display and interactive graphics. As a matter of fact, such graphics can only exert their maximal utility on a computer interface through manipulating displays. After all, medical reports still heavily rely on two-dimension graphical representation for information communication. It is therefore necessary to develop an effective visualisation technique despite limited display space.

The code used to generate the figures in this manuscript is provided as an R package in the online supplementary material.

A Details for Galbraith plot

Details for different application of the radial plots are given in [28]. It is stated that, given data $(\hat{\delta}_i, \sigma_{\hat{\delta}_i})$, $i = 1, ..., K$, a radial plot is constructed by taking as xy-coordinates:

$$x_i = 1/\sigma_{\hat{\delta}_i}, \quad y_i = (\hat{\delta}_i - \delta_0)/\sigma_{\hat{\delta}_i}$$

where δ_0 is a convenient reference value. The line $y = 0$ corresponds to $\delta = \delta_0$ and the ratio y/x, the slope of the line through origin and a point i, indicates the actual value of the estimate $\hat{\delta}_i$.

When performing subgroup analysis, it is convenient that the reference is the estimate in the full population to assess for treatment heterogeneity. A first attempt would consider then $\delta_0 = \hat{\delta}_F$, where $\hat{\delta}_F$ is the treatment effect estimate in the full population. However, as $\hat{\delta}_F$ is itself a random variable, it is better to consider also its variance. Our modification then considers the xy-coordinates:

$$x_i = 1/\sqrt{\text{Var}(\hat{\delta}_i - \hat{\delta}_F)}, \quad y_i = (\hat{\delta}_i - \hat{\delta}_F)/\sqrt{\text{Var}(\hat{\delta}_i - \hat{\delta}_F)}$$

We proceed with the details for one subgroup S, which then are applied to all subgroups $i = 1, ..., K$. Consider that there are $n_S + n_{S'} = n$ patients in each the treatment and control arm, and $\lambda = n_S/n$ is the proportion of patients in the subgroup in the population and in out trial. Further consider $\hat{\delta}_F$ and $\hat{\delta}_S$ the treatment effect estimates in the full population and subgroup, respectively. Since $\hat{\delta}_F = \lambda\hat{\delta}_S + (1-\lambda)\hat{\delta}_{S'}$, we have:

$$\begin{aligned}\text{Var}(\hat{\delta}_S - \hat{\delta}_F) &= \text{Var}(\hat{\delta}_S) + \text{Var}(\hat{\delta}_F) - 2\text{Cov}(\hat{\delta}_S, \hat{\delta}_F) = \\ &= \text{Var}(\hat{\delta}_S) + \text{Var}(\hat{\delta}_F) - 2\sqrt{\lambda}\sqrt{\text{Var}(\hat{\delta}_S)}\sqrt{\text{Var}(\hat{\delta}_F)}\end{aligned}$$

If the variances for the estimates are $\text{Var}(\hat{\delta}_F) = 2\sigma^2/n$ and $\text{Var}(\hat{\delta}_S) = 2\sigma^2/(\lambda n)$, with $\text{Var}(y) = \sigma^2$ assumed known, then:

$$\begin{aligned}\text{Var}(\hat{\delta}_S - \hat{\delta}_F) &= \text{Var}(\hat{\delta}_S) + \text{Var}(\hat{\delta}_F) - 2\text{Cov}(\hat{\delta}_S, \hat{\delta}_F) = \\ &= 2\frac{\sigma^2}{\lambda n} + 2\frac{\sigma^2}{n} - 2\text{Cov}(\hat{\delta}_S, \lambda\hat{\delta}_S + (1-\lambda)\hat{\delta}_{S'}) = \\ &= 2\frac{\sigma^2}{\lambda n} + 2\frac{\sigma^2}{n} - 2\lambda\text{Cov}(\hat{\delta}_S, \hat{\delta}_S) = \\ &= 2\frac{\sigma^2}{\lambda n} + 2\frac{\sigma^2}{n} - 4\lambda\frac{\sigma^2}{\lambda n} = 2\frac{\sigma^2}{\lambda n} - 2\frac{\sigma^2}{n} = 2\frac{(1-\lambda)}{\lambda}\frac{\sigma^2}{n}\end{aligned}$$

In the time-to-event endpoint case, we take $\hat{\delta}_S$ to be the log-hazard ratio in the subgroup S and $\hat{\delta}_F$ to be the log-hazard ratio in the full population. Consider now τ to be the number of events in the subgroup divided the total number of events, then the variance of their difference is given by:

$$\begin{aligned}\text{Var}(\hat{\delta}_S - \hat{\delta}_F) &= \text{Var}(\hat{\delta}_S) + \text{Var}(\hat{\delta}_F) - 2\text{Cov}(\hat{\delta}_S, \hat{\delta}_F) = \\ &= \text{Var}(\hat{\delta}_S) + \text{Var}(\hat{\delta}_F) - 2\sqrt{\tau}\sqrt{\text{Var}(\hat{\delta}_S)}\sqrt{\text{Var}(\hat{\delta}_F)}\end{aligned}$$

References

[1] Mohamed Alosh, Mohammad F Huque, Frank Bretz, and Ralph B D'Agostino. Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. *Statistics in medicine*, 36(8):1334–1360, 2017.

[2] Thomas Ondra, Alex Dmitrienko, Tim Friede, Alexandra Graf, Frank Miller, Nigel Stallard, and Martin Posch. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of biopharmaceutical statistics*, 26(1):99–119, 2016.

[3] Alex Dmitrienko, Christoph Muysers, Arno Fritsch, and Ilya Lipkovich. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *Journal of biopharmaceutical statistics*, 26(1):71–98, 2016.

[4] David P. Byar and Sylvan B. Green. The choice of treatment for cancer patients based on covariate information: application to prostate cancer. *Bulletin du Cancer*, 67:477–490, 1980.

[5] Patrick Royston and Willi Sauerbrei. Multivariable model-building: Advanced prostate cancer dataset, 2008. Accessed: 2017-06-01.

[6] Gerd K Rosenkranz. Exploratory subgroup analysis in clinical trials by model selection. *Biometrical Journal*, 58(5):1217–1228, 2016.

[7] Richard M Heiberger and Burt Holland. *Statistical analysis and data display: an intermediate course with examples in R*. Springer, 2015.

[8] Jonathan Swinton. Venn diagrams in r with the vennerable package. 2009.

[9] Henry Heberle, Gabriela Vaz Meirelles, Felipe R da Silva, Guilherme P Telles, and Rosane Minghim. Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics*, 16(1):169, 2015.

[10] Stirling Chow and Frank Ruskey. Towards a general solution to drawing area-proportional euler diagrams. *Electronic Notes in Theoretical Computer Science*, 134:3–18, 2005.

[11] Peter Rodgers, Jean Flower, Gem Stapleton, and John Howse. Drawing area-proportional venn-3 diagrams with convex polygons. In *Diagrams*, pages 54–68. Springer, 2010.

[12] Luana Micallef and Peter Rodgers. eulerape: drawing area-proportional 3-venn diagrams using ellipses. *PloS one*, 9(7):e101717, 2014.

[13] Jack Cuzick. Forest plots and the interpretation of subgroups. *The Lancet*, 365(9467):1308, 2005.

[14] Karin Ried. Interpreting and understanding meta-analysis graphs: a practical guide. 2006.

[15] RF Galbraith. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in medicine*, 7(8):889–894, 1988.

[16] RF Galbraith. Graphical display of estimates having differing standard errors. *Technometrics*, 30(3):271–281, 1988.

[17] Krintan A L'Abbé, Allan S Detsky, and Keith O'rourke. Meta-analysis in clinical research. *Annals of internal medicine*, 107(2):224–233, 1987.

[18] Marco Bonetti and Richard D Gelber. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–481, 2004.

[19] Marco Bonetti, Richard D Gelber, et al. A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in medicine*, 19(19):2595–2609, 2000.

[20] Patrick Royston and Willi Sauerbrei. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in medicine*, 23(16):2509–2525, 2004.

[21] Willi Sauerbrei, Patrick Royston, and Karina Zapien. Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches. *Computational statistics & data analysis*, 51(8):4054–4063, 2007.

[22] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[23] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. Upset: visualization of intersecting sets. *IEEE transactions on visualization and computer graphics*, 20(12):1983–1992, 2014.

[24] Nils Gehlenborg. *UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets*, 2017. R package version 1.3.3.

[25] UpSet - Visualizing Intersecting Sets. http://vcg.github.io/upset/. Accessed: 2018-03-19.

[26] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[27] Ravi Varadhan and Sue-Jane Wang. Standardization for subgroup analysis in randomized controlled trials. *Journal of biopharmaceutical statistics*, 24(1):154–167, 2014.

[28] Rex F Galbraith. Some applications of radial plots. *Journal of the American Statistical Association*, 89(428):1232–1242, 1994.