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.

‡These authors contributed equally to this work.

* 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 might 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.

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.

Keywords: Data visualisation, treatment effect heterogeneity, personalized medicine.

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 [1]. 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. The presence of promising results can be attributed to a small sample size or to the fact that many potential subgroups are explored, which affects the credibility level of the findings.

Subgroup analyses as investigative measures are prospective or post hoc in different settings of clinical trials. Their primary proposes can be to establish efficacy claims, subgroup discovery and/or consistency assessments across subgroups. Many researchers have proposed novel approaches and designs for different categories of subgroup analysis [2–4]. It has further received extensive attention in recent clinical research for the development of stratified medicine.

Visualization techniques, when properly used, are powerful tools to reveal data. For example, it is argued that graphics, in most cases, allow a more direct interpretation than tables [5]. There is extensive literature in good statistical graphics principles in general (e.g. [6–12]) and in the healthcare sector particularly [13–15]. However, it is also true that good graphics require careful crafting [16] and there is scope to improve when it comes to figures in clinical trial reports [17, 18].

Graphical approaches are routinely employed in subgroup analysis, typically for describing treatment effect sizes of subgroups. Such visualisations encapsulate subgroup information and boost the 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 [2]. For instance, forest plots, perhaps the most widely used graphics in the subgroup analysis setting, provide no insight into the overlap of between 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, there are several desirable characteristics for graphical approaches as initial subgroup analysis tools. Displaying sample sizes underpins the credibility level of promising and adverse findings within subgroups. It is also important to reveal overlap information, as this helps in clarifying that we look at the same signal several times and enables to focus on the subgroups that have less overlap with each other. The ability to detect treatment effect heterogeneity for subgroups should be considered as well. Moreover, it is expected that the graphic is available for a large number of subgroups to serve as a potential hypothesis generator. These characteristics can certainly constitute sensible criteria for assessment.

In this paper, we attempt to develop an effective visualization approach for subgroup analysis. Our considerations apply equally to exploratory or confirmatory settings. 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. Some of these visualisations have been already proposed for subgroup analysis before (as the forest plots), some were improved in this work, and some techniques were developed for other applications and we show how they can be applied and/or extended for the visualisation of subgroups. To facilitate the discussion, we focus on a clinical trial dataset of a treatment for prostate cancer. All graphics are performed using the R statistical software [19] and the code is publicly available as an R package for reproducibility.

Although we acknowledge that the choice of colours is a challenging task when producing graphics, we do not discuss this topic in our work, as there are other sources that already do it [20, 21]. Several of the considered plots make use of colour coding to represent the magnitude of the treatment effect across subgroups, for which we use a divergent colour palette generated by the `colorspace` R package [22].

The remainder of the paper is structured as follows: in Section 2 we describe the dataset we use for illustration and present and exploit the graphical approaches for displaying subgroup information. We differentiate among plots that allow a direct comparison of subgroup treatment effects (Section 2.1), those that provide an indirect comparison by displaying responses in treatment and control groups across subgroups (Section 2.2), and those that only allow visualizing subgroup composition or overlap between subgroups (Section 2.3). Each technique is further assessed based on a set of criteria. We summarise the assessments and features of all approaches in Section 3. We remark their practical utility and implications in clinical trials and outline potential visualisation techniques in the end.

2 Graphical approaches to subgroup problems

Our framework to assess the properties of the graphical displays consist of the following criteria:

C1 whether the plot displays effect sizes for subgroups;

C2 whether it exhibits subgroup sample sizes;

C3 whether it shows subgroups overlap information;

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

C5 whether it is applicable to a large number of total subgroups (more than 10)

We additionally point out other noticeable features of each graphical display and discuss other important characteristics, such as whether it is possible to display uncertainty or precision of the estimates and the number of subgroup-defining covariates that can be considered.

We use a prostate carcinoma dataset from a clinical trial [23] which is available on the web [24]. The data has been used before to illustrate subgroup selection methods [25]. The first publication in which this data was used dates back to 1980, where authors already discussed methodology "to determine whether comparisons of treatment in various subsets of patients yield sufficiently different results to justify the idea that there may be an optimal treatment for each patient based on his individual characteristics" [23].

The dataset consists of 475 subjects randomized to a control group or diethylstilbestrol. We are interested in identifying subgroups of patients that may benefit from the treatment. There are six covariates to consider: existence of bone metastasis (bm), disease stage (3 or 4), performance rating (pf: 0, normal; 1, limitation of activity), history of cardiovascular events (hx), age, and weight. While age and weight are continuous covariates, we categorize them for illustrative purposes in some of the plots.

As the considered endpoint is survival time in months, most of the graphical approaches we employ use the log-hazard ratio for treatment versus control as treatment effect measure. However, some of the graphics require not only a relative measure of the treatment effect but also an estimate for the response in control and treatment arms. In such cases, the difference in restricted mean survival time (RMST) is considered as the treatment effect measure.

2.1 Graphical approaches with a direct comparison of treatment effects

In this section, we devise graphics that represent or provide a measure of the treatment effect (e.g. hazard ratio) and therefore allow a direct comparison across subgroups.

2.1.1 Level plot

Level plots are typically used to show geographic surfaces in a plane. In the subgroup analysis setting, two categorical variables are arranged on the axes, and the main plot area consists 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 1a, 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 categorized into three levels (age: young=[48,65], middle-aged=(65,75], old=(75,89]; weight: low=[69,90], mid=(90,110], high=(110,152]). For each subgroup, a Cox proportional hazards model with treatment as the independent variable is fitted to obtain the estimate for the hazard ratio. Alternatively, a single multivariate model with treatment by subgroup interactions may be fitted to obtain the estimates. A divergent colour scale with a range from -3 to 3 is used to represent the log-hazard ratio. We also add the point estimate and confidence interval for the overall population in the legend as a reference. We 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.

This graphical approach is attractive since it 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 it is harmful. However, the variability of the subgroup estimates is not 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 partially. Level plots may only display the 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. 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 covariates may be arbitrary, level plots are better suited for categorical ones.

Examining Figure 1, we may conclude that the treatment is actually worse for older patients and young patients with low weight, as the direction of the treatment effect is reversed. 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 some subgroups may lead to highly variable estimates.

As a possible improvement, we propose to draw the coloured squares inside each cell with areas representing the proportion of the subgroup sample sizes relative to the full population (Figure 1b). This new design feature allows comparing 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.

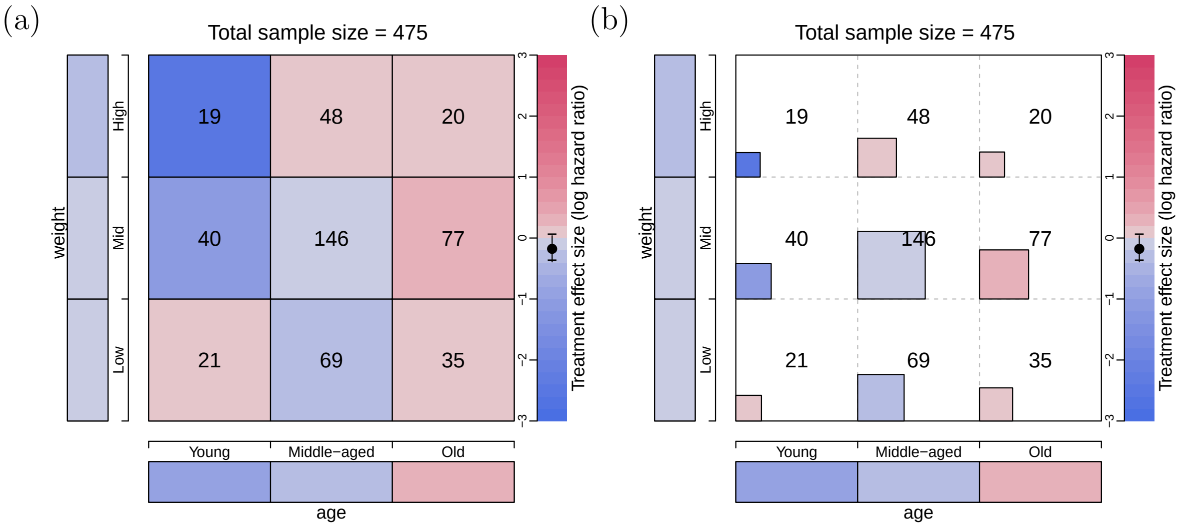


Figure 1: Level plots of treatment effect in terms of the log-hazard ratio across mutually disjoint subgroups defined by age and weight categorized in three levels. The cells on the bottom and the left margins correspond to the marginal subgroups defined by the levels of age and weight. In (b), the area of each square inside the cells is proportional to the sample sizes, which are also displayed in the middle of the cells.

2.1.2 Mosaic Plot

Mosaic plots are useful to represent contingency tables by 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 2a. Although the sample size annotation in each mosaic could be easily added, we omit it here as the sample sizes are depicted through the area of the mosaics. The interpretation of this plot is similar to the level plot presented in Figure 1b.

Mosaic plots offer the advantage that a larger number of covariates can be arranged. In Figure 2b, we use 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 rating 1, and the existence of bone metastasis. When adding additional covariates, however, it is not possible to show the information on marginal subgroups as in Figure 2a, 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 have limitation of activity (performance rating is 1).

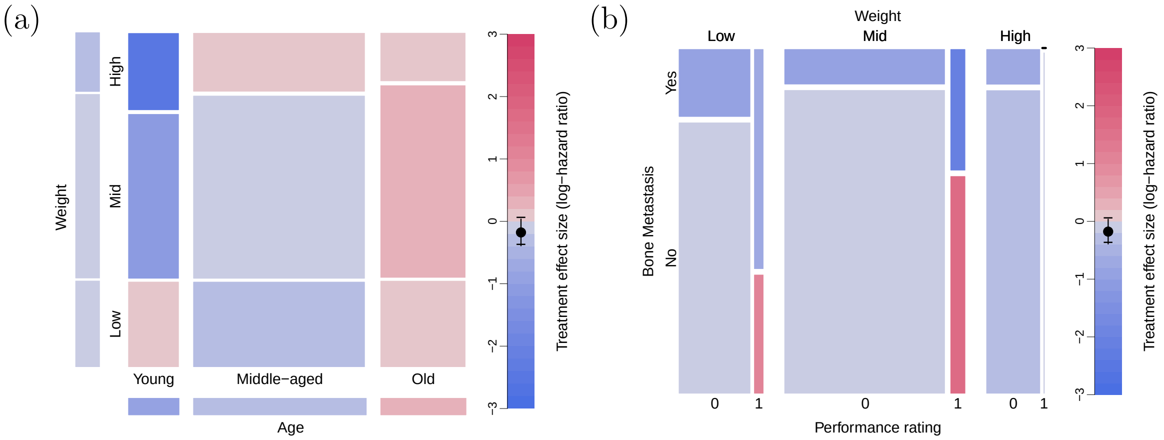


Figure 2: (a) Mosaic plot of treatment effect in terms of the log-hazard ratio across mutually disjoint subgroups defined by age and weight categorized in three levels. The cells on the bottom and the left margins correspond to the marginal subgroups defined by the levels of age and weight. The area of each mosaic is proportional to the sample sizes. (b) Mosaic plot of treatment effect in terms of the log-hazard ratio across mutually disjoint subgroups defined by weight, performance rating and bone metastasis.

2.1.3 Contour plot

An alternative to a level plot that is more suitable for illustrating continuous changes in relevant factors is a contour plot. We propose two different implementations of contour plots for the treatment effects across age and weight. In Figure 3a, we first form subgroups with a specified sample size by neighbouring subjects in terms of their values of age and weight. Subgroups of sample sizes N_{11} are formed by using a sliding window across the values of age with an overlap of N_{12} subjects. Subsequently, each subgroup is further divided into smaller subgroups of sample sizes N_{21}, using again sliding window with an overlap of N_{22}. Sample sizes and overlap to form subgroups are adopted by design based on sensible judgement. For example, subgroups should have a considerable sample size to ensure that patients in both treatment and control arms are represented. For each formed subgroup, we then calculate the log-hazard ratio for treatment versus control. The contour areas are obtained through a bivariate interpolation and smooth surface fitting (loess) for irregularly distributed data points over the range of values from the subjects under study. We also use a divergent colour scale for the effect sizes. A limitation of this approach is that there may be regions of the covariate space in which the treatment effect estimates are not reliable due to small sample sizes or no data points.

We also propose using local regression techniques to calculate the treatment effect at each coordinate. In Figure 3b, a weighted Cox proportional-hazards model is fitted at each combination of weight and age (using a step of 1 unit). A normal kernel with the centre at the coordinate values under considerations is used to assign weights to each subject. If there are less than 20 subjects within 2 standard deviations, the effect size is not calculated and the area is left blank. This helps to avoid extrapolating the results to areas in which we do not have enough information.

According to our assessment, contour plots match criteria C1 and C5 but not C2, C3 and C4. The total number of subgroups can be more than ten by controlling the overlap proportions with neighbouring subgroups. However, there is no graphical representation about subgroup sample sizes and overlap proportions.

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 covariate range. Moreover, the interpolated effect sizes may be unreliable in the regions where there are no data points or only sparse points are irregularly distributed. 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. This graphical approach only considers two continuous covariates. We also acknowledge that there may be other implementations of this plot. For example, it may be also possible to use other local regression algorithms to calculate the treatment effect at each coordinate. Recent proposals that investigate the predicted individual treatment effect can be applied to predict the effect of treatment across the covariate space [26, 27].

We observe a similar pattern to 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.

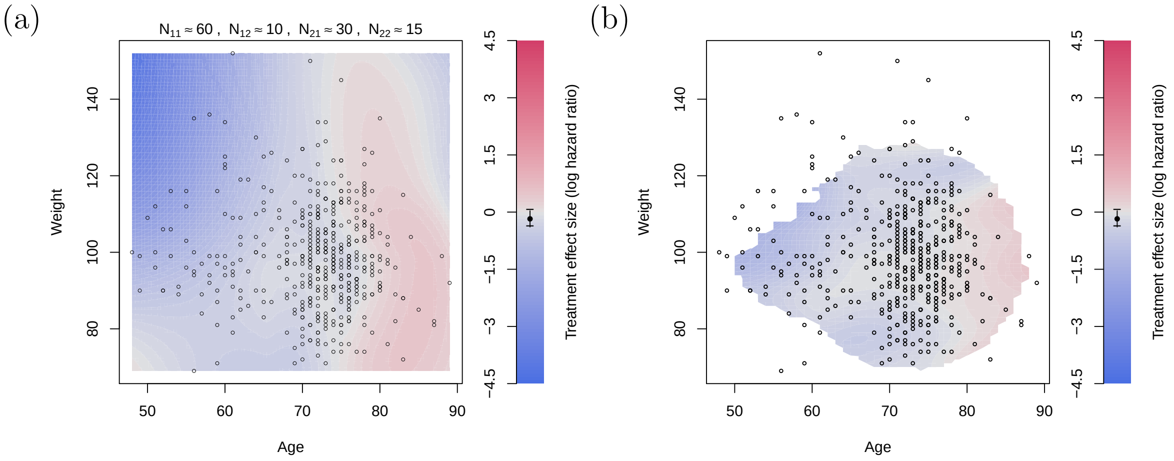


Figure 3: Contour plot of treatment effect in terms of the log-hazard ratio over the plane of age and weight. (a) Contour lines are drawn by forming subgroups with neighbouring subjects, calculating the treatment effect for subgroups and interpolating the results using loess. N_{11} stands for the sample 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. (b) Contour lines are drawn by fitting a local regression at each point of the grid, using subjects weights according to their distance to the point of the grid. Points with few subjects in the vicinity of the grid point were left blank

2.1.4 Venn diagram

Venn diagrams are undoubtedly the most widely used tool to visualize sets and their relations. In the subgroup analysis setting, Venn diagrams may be used 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 4a. Each circle defines the subgroup of patients for which the level of the corresponding variable is ”yes” or 1. 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.

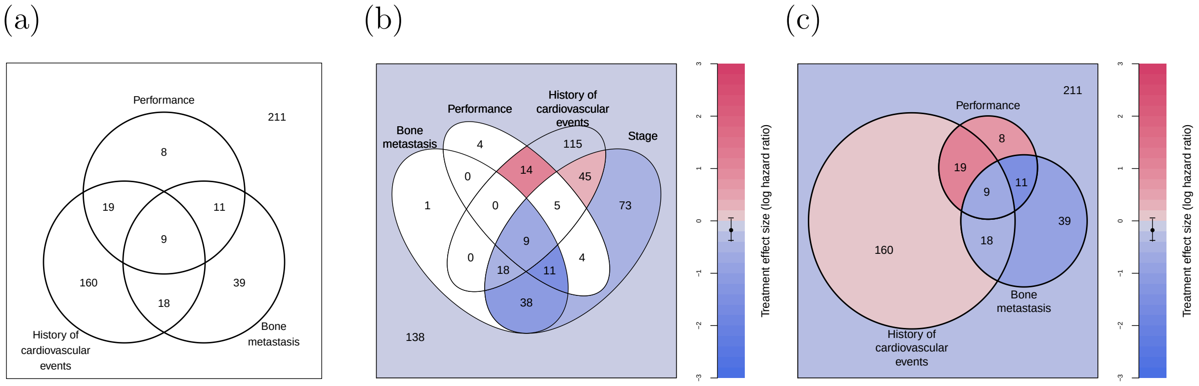


Figure 4: (a) Venn diagram of 3 subgroups defined by presence of bone metastasis, history of cardiovascular events, and performance rating = 1. (b) Venn diagram of 4 sets defined by presence of bone metastasis, disease stage, performance rating = 1 and history of cardiovascular events with treatment effect sizes in terms of the log-hazard ratios. (c) Approximate area-proportional Venn diagram of 3 subgroups defined by presence bone metastasis, history of cardiovascular events and performance rating = 1 with treatment effect sizes in terms of the log-hazard ratios.

Figure 4b and 4c consider Venn diagrams with four and three subgroups, respectively. Both represent the treatment effect in terms of the log-hazard ratio by colouring the corresponding regions. This feature thus enables Venn diagram to satisfy the criterion C1. However, it does not serve for detecting heterogeneity in subgroup effect sizes because variability of the estimates is not given.

As seen in Figure 4b, using four ellipses for representing all possible subgroups (formed through intersection and complement) is visually appropriate. Other shapes (such as polygons [28, 29]) 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 4c further considers proportional-area 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 employ the simple algorithm mentioned in [30]. In fact, other algorithms to display each region area proportional the sample sizes are available. Recently an algorithm that can produce accurate area-proportional Venn diagrams using ellipses was developed [30]. However, the algorithm is somewhat sophisticated and can only work on three sets.

Venn diagrams satisfy C2, C3 in our assessment. Useful extensions to Venn diagrams, such as the Edwards' construction [31, 32], 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.

Figure 4 shows that the treatment effect is reversed, with the control treatment being better than the experimental one for those subjects without bone metastasis when they have previous cardiovascular events or limitation of activity (performance rating is 1).

2.1.5 Bar chart

Another useful graphical technique to depict treatment effect sizes is 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 categorized age and weight variables as in previous examples. In Figure 5. each covariate is categorized into three levels and the bars represent mutually disjoint subgroups. The levels of age and weight are respectively listed at the top and the bottom part of the picture. The height of the bars is proportional to the treatment effect differences between the treatment/control arms, that is, the difference in RMST. The width of the bars is proportional to the subgroup sample sizes. This arrangement has, therefore, 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 show which subgroups have 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 covariate is larger.

Few noteworthy characteristics also need to be mentioned. First, 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. Second, although it satisfies C5, a high number of covariates or levels may be problematic, making it difficult to compare the widths of the bars. Third, as in level plots, the cut-off points for categories in continuous variables may be arbitrary and categorical covariates are therefore preferred for bar plots.

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.

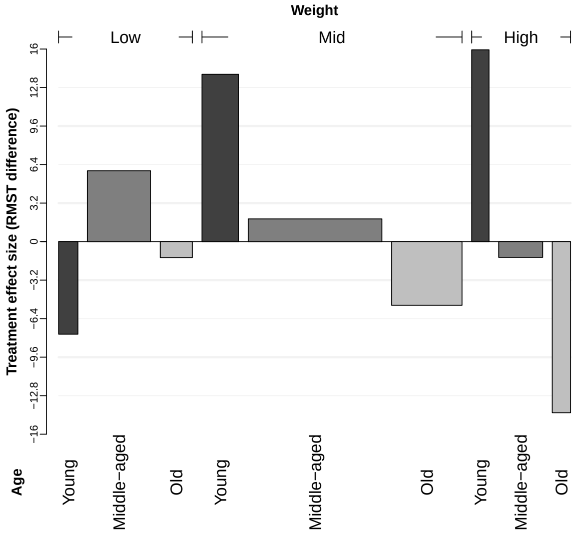


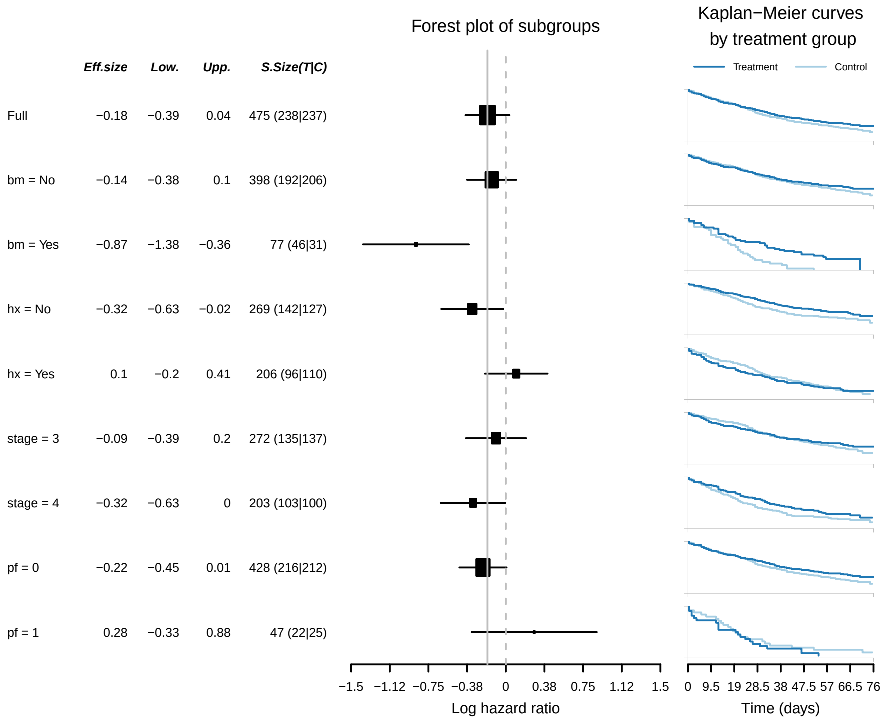
Figure 5: Bar plot of treatment effect in terms of the difference in restricted mean survival time across mutually disjoint subgroups defined by age and weight categorized in three levels. The width of each bar is proportional to the sample size for subgroups. The area can be interpreted as the gain/loss in restricted mean survival when using treatment in comparison to control. The black, grey and light grey colours indicate the age categories young, middle-aged and old, respectively.

2.1.6 Forest plot

Although forest plots are a common graphical display approach for meta-analysis [33], they are also extensively used for subgroup analysis [34, 35]. 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 [34]. Figure 6 shows its application for the prostate cancer dataset considering four binary covariates. The main panel in the middle displays the subgroup treatment effects with their confidence intervals. The squares in the centre of each error bar are proportional to the subgroup sample sizes. Additional information in a table format is usually included to provide the exact magnitude of the estimates. 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). When using a continuous or binary endpoint, it is also recommended to include the 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 meet all the criteria but C3 because of the inability to show subgroup overlaps.

We observe in Figure 6 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.



	Eff.size	Low.	Upp.	S.Size(T|C)
Full	−0.18	−0.39	0.04	475 (238|237)
bm = No	−0.14	−0.38	0.1	398 (192|206)
bm = Yes	−0.87	−1.38	−0.36	77 (46|31)
hx = No	−0.32	−0.63	−0.02	269 (142|127)
hx = Yes	0.1	−0.2	0.41	206 (96|110)
stage = 3	−0.09	−0.39	0.2	272 (135|137)
stage = 4	−0.32	−0.63	0	203 (103|100)
pf = 0	−0.22	−0.45	0.01	428 (216|212)
pf = 1	0.28	−0.33	0.88	47 (22|25)

Figure 6: 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.

2.1.7 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 7 shows a tree plot of treatment effect differences for subgroups defined by bone metastasis, performance rating and history of cardiovascular events. In each level or 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 subgroup sample sizes. An additional horizontal dotted line is added at each level for the overall treatment effect size. In Figure 7a, the y-axis for each level of the plot is drawn independently from the others levels. In the Figure 7b, the y-axes are consistent across levels, which helps to visualize the difference in variability of the estimates.

Tree plots match all the criteria. It is obviously fit to C1, C2 by design. In addition, the subgroups at 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 subgroups 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 the number of covariates that are involved and how many categories each covariate has.

A few features of tree plots are worthily pointed out. First, it provides information on 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 7 allows us to draw additional conclusions regarding the treatment effect sizes. We 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 rating 1 has a positive log-hazard ratio, implying that the control is better than treatment for this subgroup.

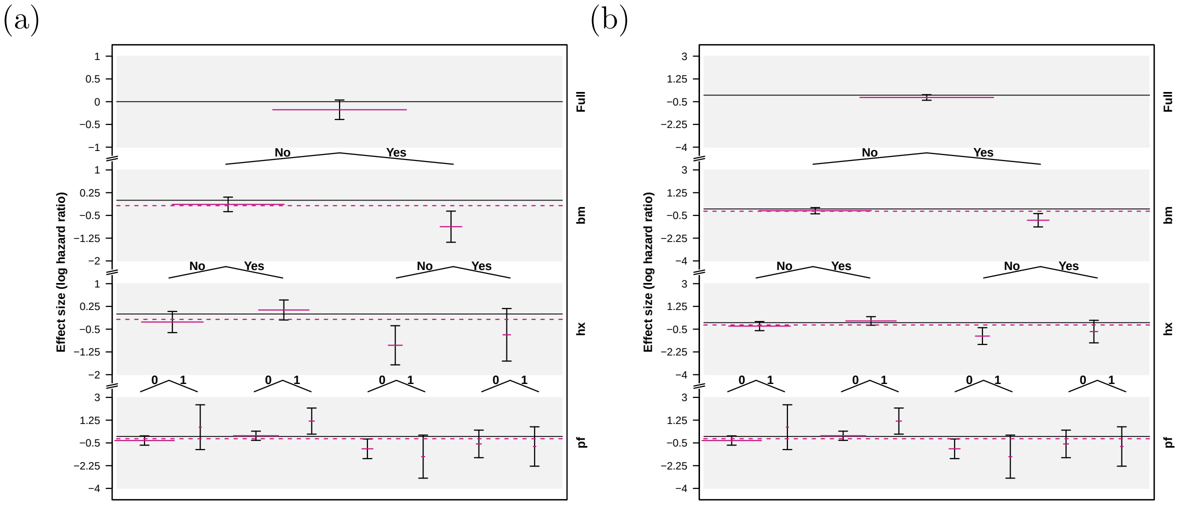


Figure 7: Tree plot for the treatment effect in terms of the log-hazard ratio for subgroups defined by category combinations of existence of bone metastasis (bm), history of cardiovascular events (hx), and performance rating (pf). Each layer shows the 95% C.I. of treatment effect differences for the 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. In (a) the y-axes are independent in each layer of the plot, while in (b) y-axes are kept fixed across levels, which allows comparing variability in the estimates

2.1.8 Galbraith plot

A Galbraith plot [36, 37] is an alternative or supplementary to a forest plot for examining heterogeneity of studies or subgroups in a meta-analysis. The variant that is shown in Figure 8 exhibits the estimation of treatment effect sizes for subgroups defined by the four binary covariates. The xy-coordinates correspond to the points:

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

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 treatment effect heterogeneity if one standardized estimate is located outside the band. The slope of the line from the origin through each subgroup point corresponds to the effect size estimate $\hat{\delta}_i$ of the corresponding subgroup. An additional radial axis is drawn to depict the subgroup effect sizes, which are represented with the red tick marks. The central line at $y = 0$ points to the average treatment effect for the full population.

We note here that, as $\hat{\delta}_F$ is itself a random variable, it may better to consider also its variance. An additional modification may then consider the xy-coordinates:

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

The resulting plot using such these values is given in the Appendix A. The drawback of this modification is that the x-axis does no longer represent the standard error of the treatment effect estimate.

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 of subgroups and is also helpful to detect outliers. Galbraith plots only partially fit 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 does not display 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.

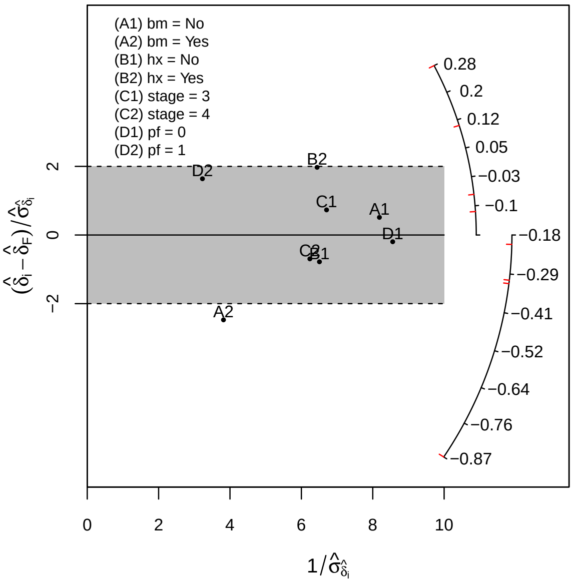


Figure 8: Galbraith plot for subgroups defined by existence of bone metastasis (bm), history of cardiovascular events (hx), stage, and performance rating (pf).

2.1.9 L'Abbé plot

L'Abbé plots [38] are a variant of scatter plots which are useful for examining heterogeneity in a 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 9). the $x-$ and $y-$coordinate for each subgroup 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 and treatment arms in the subgroup. We draw a diagonal dashed line at $y = x$ which 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 magnitude of the effect size, that is 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.

L'Abbé plots share the same graphical assessment results as 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 an additional conclusion in our example. The subjects with bone metastasis in the control group have a lower RMST. When receiving the experimental treatment, however, the RMST is closer to that of the other subgroups.

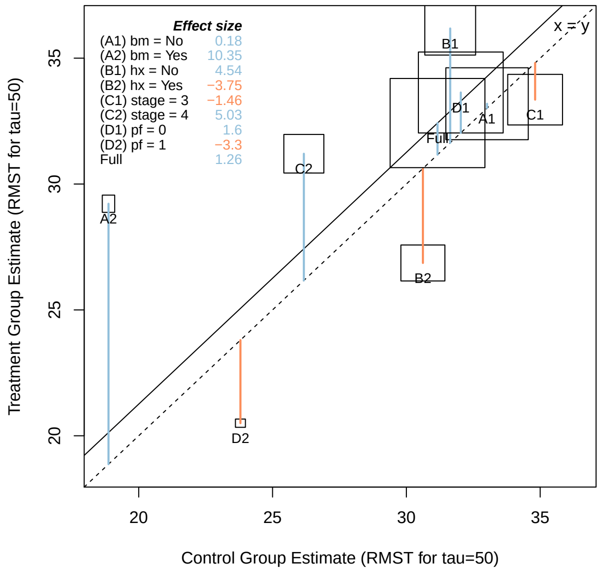


Figure 9: L'Abbé plot for subgroups defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes are given in terms of the difference in restricted mean survival time (RMST).

2.1.10 STEPP

The subpopulation treatment effect pattern plot (STEPP) [39, 40] gained popularity in breast cancer recently. It is a non-parametric method mainly for examining whether treatment-covariate interactions exist. In Figure 10, 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% overlap with 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 worth noting that the point estimates (mean, boundaries by 95% simultaneous C.I. and individual C.I.) are positioned at the mean value of age for each subgroup. If the green line does not lie in the region formed by simultaneous confidence intervals, it reveals that interaction may exist.

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 [41, 42] or non-parametric methods (such as Gaussian processes [43]), 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 fluctuates around the overall treatment effect. When approaching the ends of the range of the covariate the estimate of the log-hazard ratio departs from the estimate for the full population, although the confidence intervals still cover it.

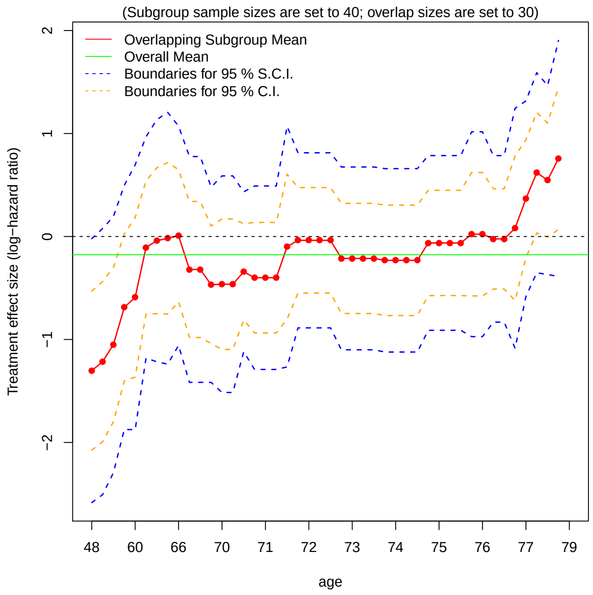


Figure 10: STEPP plot of overlapping subgroups defined by age. Each subgroup has a sample size of around 40 (N11 = 40) and is controlled to have about 87% (N12/N11) being overlapped with the neighbouring subgroups.

2.1.11 UpSet Plot

UpSet plots are a novel visualization technique for the quantitative analysis of sets and their intersections [44]. 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 11, we use the `UpSetR` R package [45] to create the plot with the six subgroup-defining covariates. The variable age is dichotomized (1: >75 years), as is weight (1: >100) to have binary covariates. 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 on the bottom allows visualizing the composition of the subgroup by showing which sets are intersected. The main bar plot displays the sizes of the subgroups that are defined by the respective intersections. For example. the first and tallest bar indicates there are 52 subjects with performance rating 0 (normal), no existence of bone metastases, age ≤ 75, weight > 100, disease stage 3, and no history of cardiovascular events. Moreover, we add a 'query' to display the frequency of treatment and control in each subset.

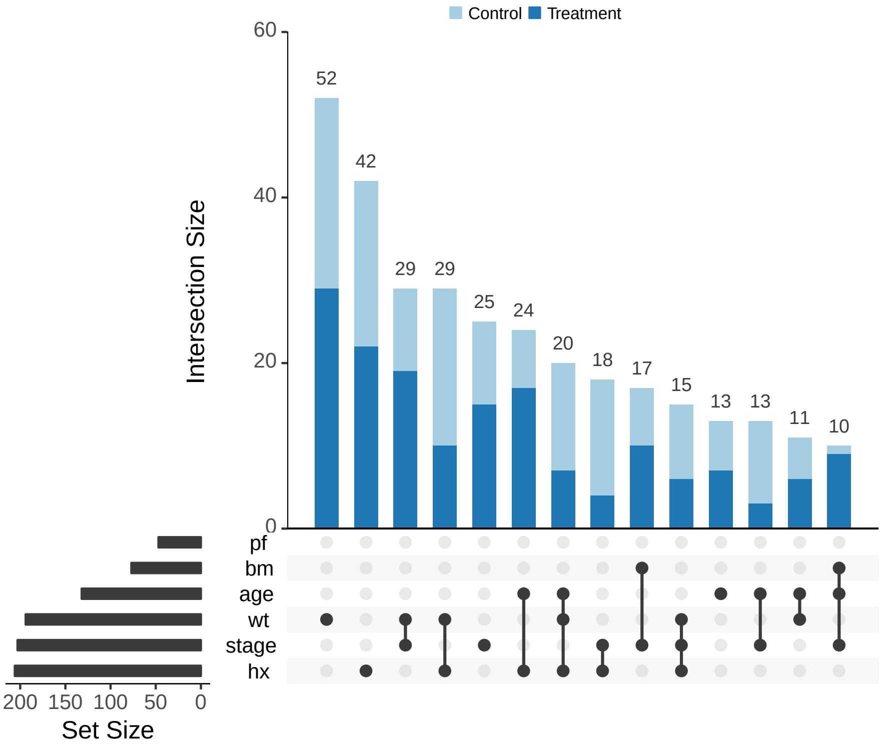


Figure 11: Upset plot displaying the subgroups conformed by the intersection of all subgroup-defining covariates

We extend the `UpSetR` R package to display effect sizes in an extra panel (Figure 12). In this case, the log-hazard ratio and its confidence interval are shown. This information is similar to 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 displaying lower level intersections. We implement a new icon for the matrix panel: a '+' symbol if a variable is equal to 1 or 'yes', a '-' if a 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 overall population (no subgroup division), which has a size of 475. The second bar with a size of 428 corresponds to the subgroup of normal performance rating (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 of subgroups according to different criteria, such as total subgroup sample size, sample size per treatment, etc. Finally, the bar plot on top of the matrix panel indicates the marginal set sizes in relation to the total sample size, with the black region corresponding to the 1 or 'yes' category and the white region corresponding to the 0 or 'no' category.

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. 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.

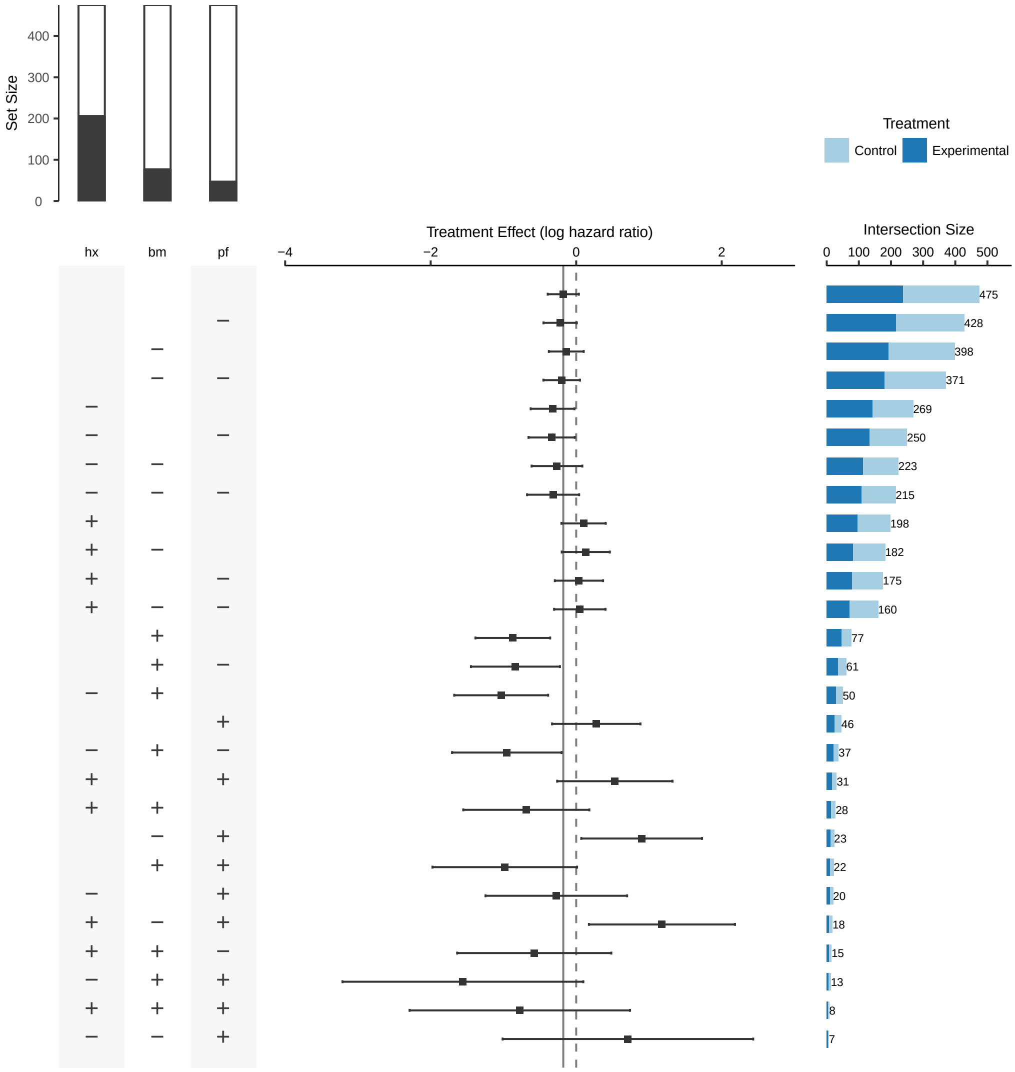


Figure 12: Improved UpSet plot for subgroups defined by performance (pf), bone metastasis (bm) and history of cardiovascular events (hx). The panel on the left (matrix) displays how the subgroups are formed by assigning a '+' if the variable is equal to 1 and a '−' if the variable is equal to 0. The bar plot on top of the matrix panel indicates the marginal set sizes in relation to the total sample size, with the black region corresponding to the 1 or 'yes' category and the white region corresponding to the 0 or 'no' category. Treatment effect sizes and their confidence intervals are display in the panel on the middle and the subgroup sizes in the horizontal bar plot on the right.

2.1.12 Circle Plot

Circular diagrams are widely used to visualize genomic data [46]. There are several approaches for the use of these diagrams, although the main aspect is that it allows representing the relationships between pairs of sets. For our example, we use the categorized variables age and weight (Figure 13). The categories of each variable are arranged along the circle, where each of their corresponding cells has a size proportional to the corresponding subgroup 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 relative overlap between the categories of the variables. Their width is calculated in correspondence to the proportion of subjects from a subgroup that is also in the subgroup to which the bands connect.

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, especially when the number of covariates is extremely large as when dealing with genomic data.

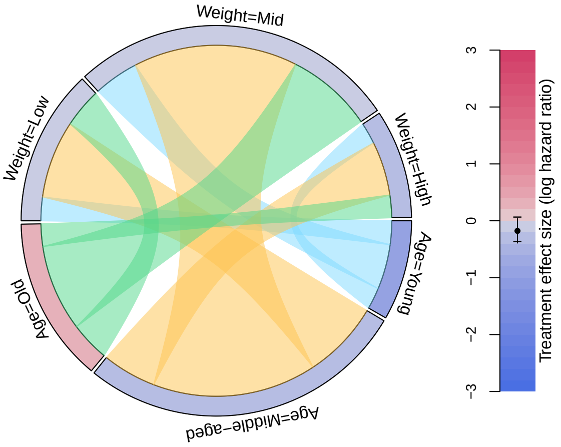


Figure 13: Circle plot for the subgroups formed by age and weight. The colours along the circle represent the treatment effect in terms of the log-hazard ratio. The ribbons that link the subgroups represent their overlap.

2.2 Graphical approaches with an indirect comparison of treatment effects

In some cases, it may be also of interest to visualize responses by treatment arm across subgroups. For example, we may want to display the survival or mortality rate, or simply the mean response if a continuous endpoint is considered. The following plots that we consider are examples of graphics that allow an indirect comparison of treatment effects.

2.2.1 Mosaic Plot

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 [11]. Although this plot may be more appropriate when the endpoint is binary, it is possible to adapt it for survival endpoint. In Figure 14, we use 2-year survival (blue corresponds to 'yes') by treatment and age category. 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, indicating that the treatment effect may not be homogeneous across the levels of age.

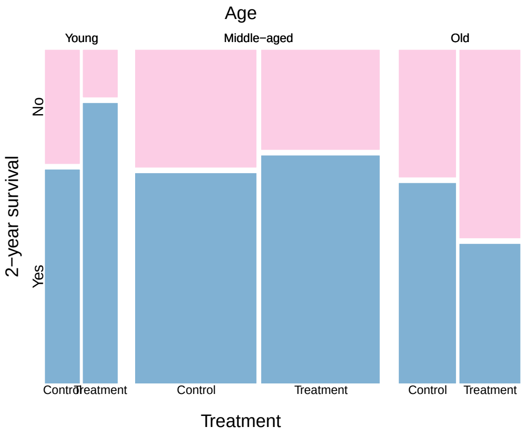


Figure 14: Mosaic plot displaying 2-year survival by treatment arm for the subgroups formed by age categories.

2.2.2 Coxcomb plot (Nightingale rose)

A Nightingale coxcomb plot [47] is a type of radial plot that was introduced in 1858 and is usually recommended as an alternative to pie charts [9]. In Figure 15, we arrange the subgroups defined by the categorized age and weight variables along the circle using a combination of bar plot and polar coordinates with the ggplot2 [48] R package. In this plot, the angles that define each sector are kept fixed, but the angles vary proportionally to the square root of the sample size in each subgroup to perceive areas adequately. We colour the areas according to the 2-year survival rate of each subgroup. In Figure 16 we further divide the plot into treatment and controls arms. This feature allows to check the sample sizes per subgroup by arm and may help visualize differences in the survival rates.

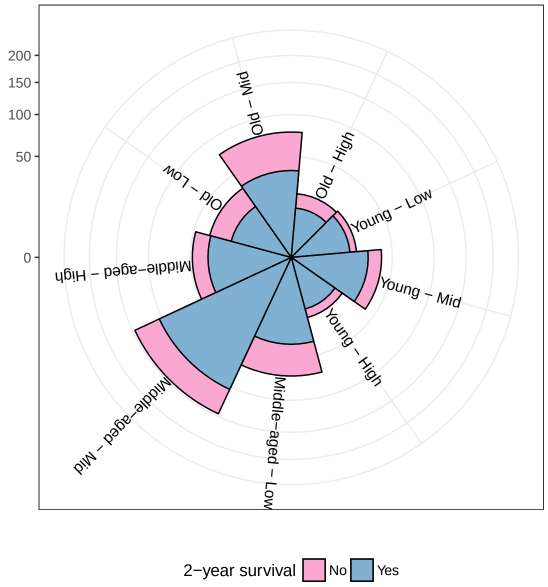


Figure 15: Nightingale coxcombs plot for subgroups defined by age and weight with 2-year survival rate. The radius of the sectors are proportional to the square root of the sample sizes in the subgroups

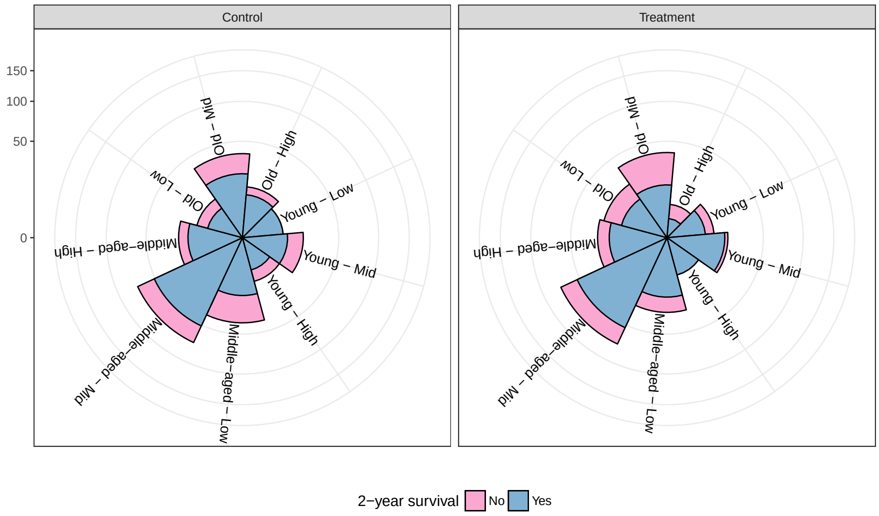


Figure 16: Nightingale coxcombs plot for subgroups defined by age and weight with 2-year survival rate and separated by treatment arm. The radius of the sectors are proportional to the square root of the sample sizes in the subgroups

2.2.3 Alluvial diagram

Alluvial diagrams are flow diagrams that can be used to display the distribution of the subjects across the subgroup-defining covariates. 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 17 shows one possible implementation of the alluvial plot for the 2-year survival per treatment arm across levels of performance, history of cardiovascular events and bone metastasis. The plot was generated using the `alluvial` R package [49].

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.

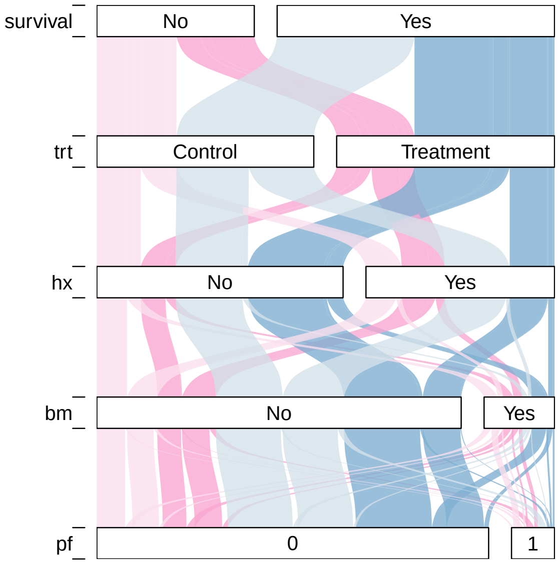


Figure 17: Alluvial diagram displaying the distribution of patients across the subgroups defined by history of cardiovascular events (hx), existence of bone metastasis (bm) and performance rating (pf). The dark bands correspond to patients that were randomized to treatment while lighter ones to patients in control. Blue coloured bands represent patients that had survived for at least 2 years, while pink ones represent those who did not. The width of the bands is proportional to the sizes of the subgroups.

2.3 Graphical approaches for subgroup composition

Figure 18 shows another implementation of alluvial plots for subgroup composition. 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 alluvium (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).

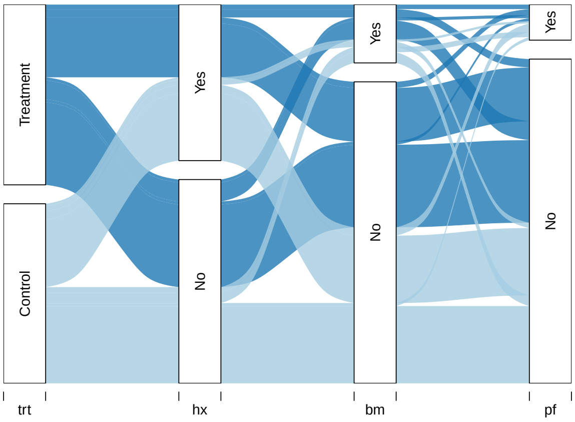


Figure 18: Alluvial diagram displaying the distribution of patients across the subgroups defined by history of cardiovascular events (hx), existence of bone metastasis (bm) and performance rating (pf). The dark blue bands correspond to patients that were randomized to treatment while light blue ones to patients in control. The width of the bands is proportional to the sizes of the subgroups.

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 that are shown in Figure 19 exhibit certain subgroup information about pairwise overlap proportions or similarity measures. Figures 19a- 19d show pairwise relative overlap proportions, where different colours show the range of overlap magnitude.

More specifically, Figure 19a 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 19b is a variant of Figure 19a. 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 19c is a plot merely using coloured lines connecting subgroup labels on different levels. This plot should be read from top to bottom. A subgroup label on the higher level is the baseline subgroup for the relative overlapping proportions with its counterpart on the lower level. Figure 19d 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 19e-19f show dissimilarity distance, which is defined by one minus a relative overlap proportion. Each line of Figure 19e 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 19f shows the same information as Figure 19e. Note that for each subgroup we do not show its dissimilarity distance to itself and its complement.

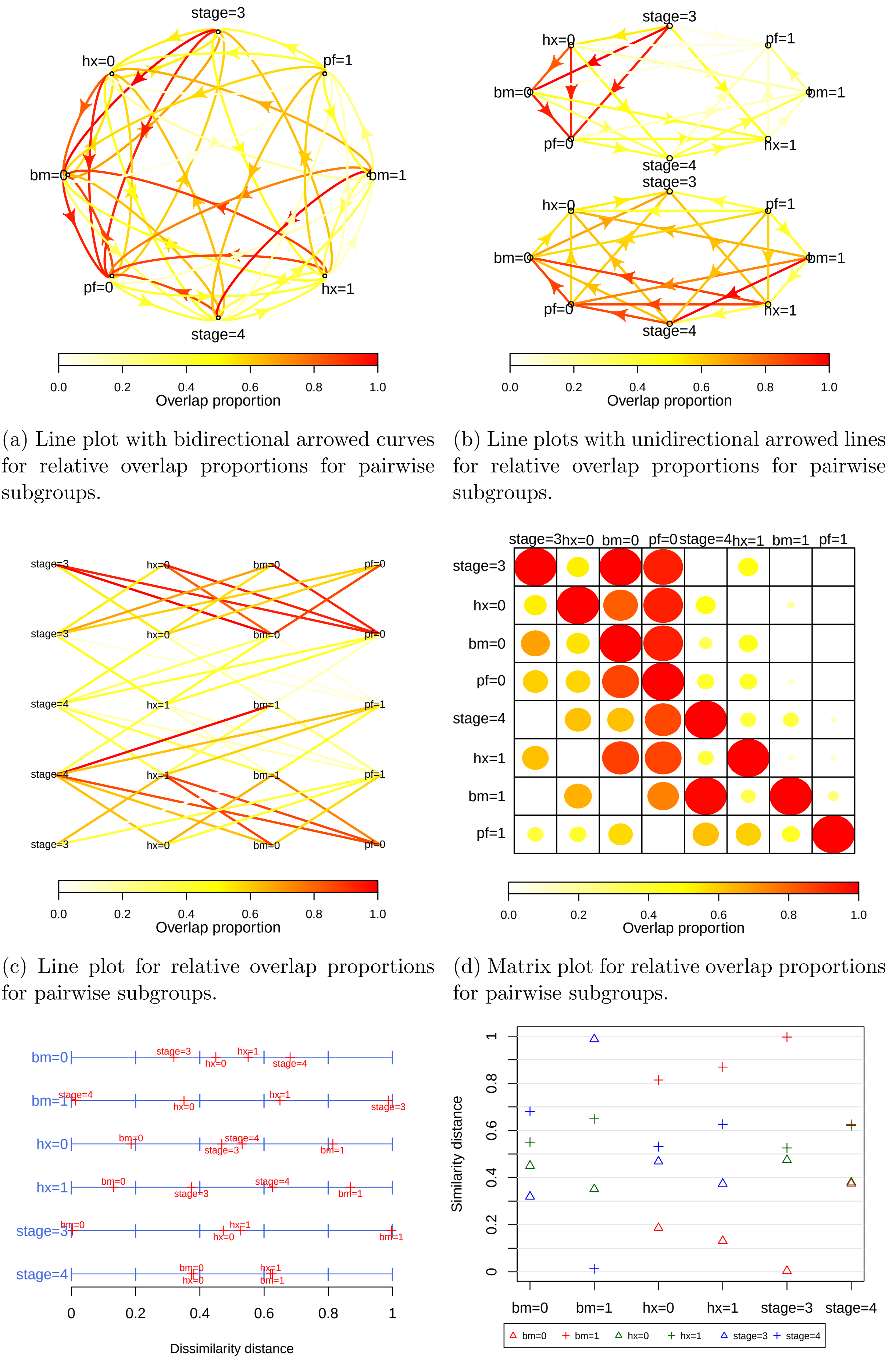
Incidentally, the Jaccard index, namely $|A \cap B|/|A \cup B|$ for any sets A, B, can replace pairwise overlap proportions for subgroup overlap information. The graphical display is thus simplified due to not showing repetitive Jaccard indexes. However, this measure may lead to missing some information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

In Appendix B, we present additional alternatives for the line and circle plots that display the overlap between the subgroups using a matrix layout. These plots may be easier to interpret as the plot is not overloaded with information, one can focus on one subgroup at the time. However, these plots may be impractical when having a large number of subgroups.

3 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 of displaying subgroup effect sizes. However, only a handful of them displays or has information to construct confidence intervals. In terms of the second criterion, the majority of the approaches provide a visual display on subgroup sample sizes. Galbraith plot indirectly shows the information through the standard error of estimators. Furthermore, only forest plot and L'abbé plot provide subgroup responses for the treatment and control arms. The third criterion is fully or partially hold for all apart from



(a) Line plot 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 plot for relative overlap proportions for pairwise subgroups.

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

(e) Dissimilarity measures for marginal subgroups.

(f) Dot plot for dissimilarity measures for marginal subgroups.

Figure 19: 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 subgroup overlap information; **C4**: whether to serve for detecting heterogeneity in subgroup effect sizes (or the treatment-covariate interaction); **C5**: whether is available for a 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 are provided in the plot or at least 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

	Criterion					Additional features			
	C1	C2	C3	C4	C5	C.I.	T/C Effect	Overlap	N_c
Level plot	✓	✓	✓		✓			P	2
Mosaic plot	✓	✓	✓		✓			P	2-4
Contour plot	✓		✓		✓			P	2
Venn diagram*	✓	✓	✓		✓			A	2-6
Bar chart	✓	✓	✓		✓	✓		P	1-5
Tree plot	✓	✓	✓	✓	✓	✓		A	1-5
Forest plot*	✓	✓		✓	✓	✓	✓	P	1-40
Galbraith 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

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. It is noted that when the number of subgroups is small (say, up to five), the forest plot, Galbraith plot and L'Abbé plot can combine a Venn diagram or other plots for displaying overlap of subgroups.

The capacity for detecting heterogeneity or interaction is not equipped all approaches. We issue warning here since that visualisations that do not adequately demonstrate the uncertainty of the estimates may be misleading and can lead to an over-interpretation of the heterogeneity of the treatment effect across subgroups. Some of the plots include featuring a reference line corresponding to the overall effect size. The 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 may be available to handle more than ten subgroups. However, we found that forest plots, Galbraith plots, L'Abbé plots, UpSet, and circle plots may be the best graphical approaches to handle a large number of subgroups. Venn diagrams and tree plots can practically deal with only up to five sets for effective visualisation.

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 dataset is large and the distributions of two covariates considered are roughly uniform; level plots and bar charts may be easier to understand 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 the 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, for example, a standardization technique ([50]) 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 three-dimension graphical displays 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.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567. Prof Jaki's contribution is independent research arising in part from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. Funding for this work was also provided by the Medical Research Council (MR/M005755/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

A Alternative Galbraith plot

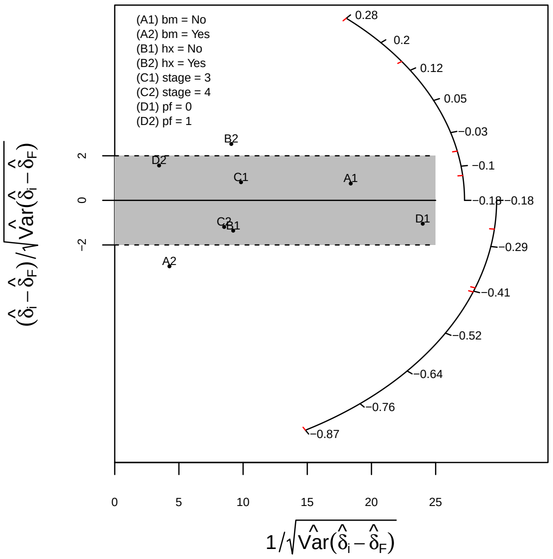


Figure 20: Modified Galbraith plot across subgroups defined by stage, history of cardiovascular events (hx) and existence of bone metastasis (bm)

B Circle plots in matrix layout

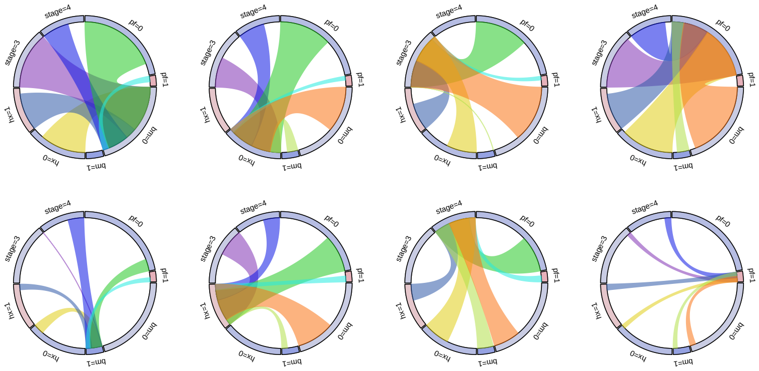


Figure 21: Circle plots displaying the subgroups conformed by the categorized age and weight covariates. The width of each section is proportional to the sample size in the corresponding subgroup. Each circle displays the relative overlap of one subgroup with the others.

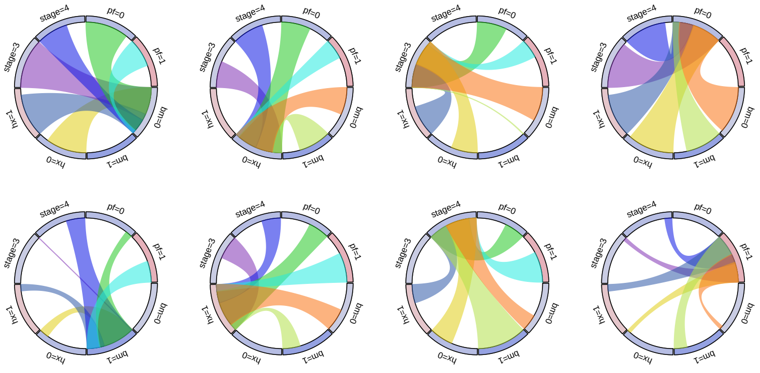


Figure 22: Circle plots displaying the subgroups conformed by the categorized age and weight covariates. Sample sizes are not depicted in this version as the widths are the same for each section. Each circle displays the relative overlap of one subgroup with the others.

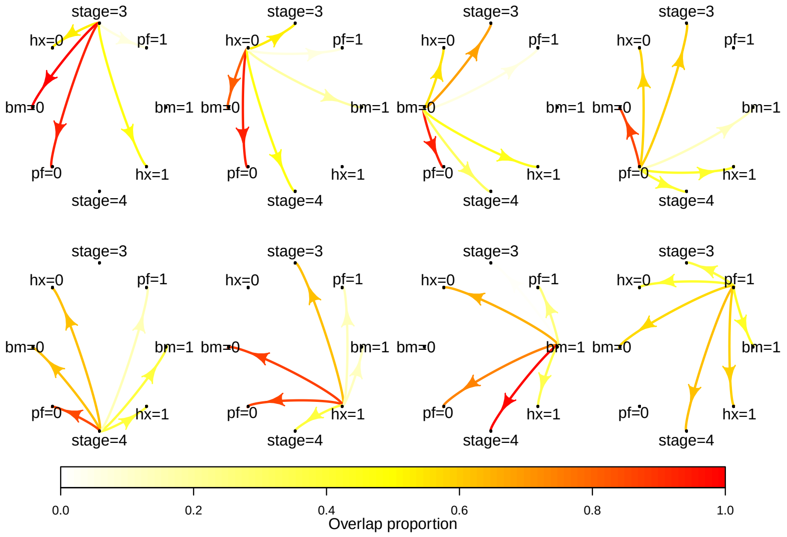


Figure 23: Line plots with unidirectional arrowed lines for relative overlap proportions for pairwise subgroups. Each subplot contains the overlap of one subgroup with all the others.

References

[1] Committee for Medicinal Products for Human Use. Guideline on the investigation of subgroups in confirmatory clinical trials. *London: European Medicines Agency*, 2014.

[2] 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.

[3] 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.

[4] 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.

[5] Andrew Gelman, Cristian Pasarica, and Rahul Dodhia. Let's practice what we preach: turning tables into graphs. *The American Statistician*, 56(2):121–130, 2002.

[6] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1983.

[7] William S Cleveland. A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*, 2(4):323–343, 1993.

[8] John W Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

[9] Naomi B Robbins. *Creating more effective graphs*. Wiley, 2012.

[10] Leland Wilkinson. *The grammar of graphics*. Springer Science & Business Media, 2006.

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

[12] Winston Chang. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. " O'Reilly Media, Inc.", 2012.

[13] Milo A Puhan, Gerben Ter Riet, Klaus Eichler, Johann Steurer, and Lucas M Bachmann. More medical journals should inform their contributors about three key principles of graph construction. *Journal of clinical epidemiology*, 59(10):1017–e1, 2006.

[14] Andreas Krause and Michael OConnell. *A picture is worth a thousand tables: graphics in life sciences*. Springer Science & Business Media, 2012.

[15] Susan P Duke, Fabrice Bancken, Brenda Crowe, Mat Soukup, Taxiarchis Botsis, and Richard Forshee. Seeing is believing: good graphic design principles for medical research. *Statistics in medicine*, 34(22):3040–3059, 2015.

[16] Martin Krzywinski. Points of view: elements of visual style, 2013.

[17] Stuart J Pocock, Thomas G Travison, and Lisa M Wruck. Figures in clinical trial reports: current practice & scope for improvement. *Trials*, 8(1):36, 2007.

[18] Jennifer C Chen, Richelle J Cooper, Michael E McMullen, and David L Schriger. Graph quality in top medical journals. *Annals of emergency medicine*, 69(4):453–461, 2017.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[20] Achim Zeileis, Kurt Hornik, and Paul Murrell. Escaping rgbland: selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270, 2009.

[21] Mark Harrower and Cynthia A Brewer. Colorbrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[22] Ross Ihaka, Paul Murrell, Kurt Hornik, Jason C. Fisher, and Achim Zeileis. *colorspace: Color Space Manipulation*, 2016. R package version 1.3-2.

[23] 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.

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

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

[26] Andrea Lamont, Michael D Lyons, Thomas Jaki, Elizabeth Stuart, Daniel J Feaster, Kukatharmini Tharmaratnam, Daniel Oberski, Hemant Ishwaran, Dawn K Wilson, and M Lee Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in medical research*, page 0962280215623981.

[27] Patrick M Schnell, Qi Tang, Walter W Offen, and Bradley P Carlin. A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, 72(4):1026–1036, 2016.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2009.

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

[35] Doron Aronson. Subgroup analyses with special reference to the effect of antiplatelet agents in acute coronary syndromes. *Thrombosis and haemostasis*, 112(01):16–25, 2014.

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

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

[38] 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.

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

[40] Marco Bonetti and Richard D Gelber. 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.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] Florence Nightingale. *Notes on matters affecting the health, efficiency, and hospital administration of the British army: founded chiefly on the experience of the late war.* Harrison and Sons, 1858.

[48] Hadley Wickham. *ggplot2: elegant graphics for data analysis.* Springer, 2016.

[49] Michal Bojanowski and Robin Edwards. *alluvial: R Package for Creating Alluvial Diagrams*, 2016. R package version: 0.1-2.

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