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 analysis are a routine part of clinical trials, for example 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 are routinely employed in subgroup analyses to depict effect sizes of subgroups and aid identification of groups that respond differentially. Such visualisations aim to encapsulate all relevant information about a subgroup and seeks to aid the clinical decision making process. However, many existing approaches do not capture all the core information and/or are prone to lead to misinterpretation of subgroup effects. In this paper we critically appraise existing visualization techniques and propose useful extensions to increase their utility

Keywords: Data visualisation, subgroup analysis, forest plot

1 Introduction

Investigating target populations potentially beneficial to an innovative intervention is essential in clinical trials. Such investigations are challengeable because various issues are needed to address. For example, one is a wide search range. Enrolling patients have rather diverse baseline characteristics for considerations, such as age, gender, race, disease severity or biomarker profiles. Another is decision making about populations for treatment use. Even if efficacy is established in the overall population, a complete benefit/risk assessments of subgroups should be undertaken before deciding the treatment to the whole or the population excluding certain subgroups. Also, the credibility level of findings is concerned. The presence of promising results can be attributed to a small sample size.

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. They are 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 effect sizes of subgroups. Such visualisation encapsulates subgroup information and boosts clinical decision making process. However, not much attention has been paid to how to make effective graphics. Existing approaches still have inherent drawbacks and their use may lead to misinterpretations to subgroup effect sizes[1]. For instance, forest plots provide no insight on the overlap of different subgroups; additionally, whether or not a subgroups confidence interval crosses the no-effect point does not necessarily imply a lack of effect or contribute an effect to the subgroup. It is therefore crucial to correctly depict effect sizes and essential information of subgroups.

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 subgroups to serve potential hypothesis generating. These characteristics can certainly constitute sensible criteria for assessments.

In this paper we attempt to develop an effective visualization approach with desired features and particularly a two-dimension display. Our considerations of developing approaches is not constrained regardless of exploratory or confirmatory settings. Also, to facilitate development we focus on a clinical trial dataset of a treatment for prostate cancer. 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.

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 nine graphical approaches for displaying subgroup information. Each technique is further assessed based on a set of criteria. Section 4 focuses on improved graphical displays and alternatives. Some are further improved by mitigating their original demerits. In Section 5 we summarise the assessment and features of all the improved approaches. Remarks on their practical usefulness and implications in clinical trials are made. We outline the 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. The p-value of the log-rank test for the test of the difference in survival between treatment and control was 0.103.

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.

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 graphical displays is depicted at first 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 to display effect sizes for subgroups;

C2 whether to exhibit subgroup sample sizes;

C3 whether to show all overlap information for subgroups;

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

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

3.1 Level plot

Figure 1 shows the application of level plots (LP) for treatment effect differences in subgroups defined by age and weight. Each covariate is partitioned into three levels and the cells represent mutually disjoint subgroups. Each subgroup is conformed by the pairwise overlap of marginal subgroups defined by the categories of age and weight. The figures inside the cells stand for the corresponding subgroup sample sizes. The three cells on the bottom and the left margins represent the marginal subgroups corresponding to the three levels of age and weight, respectively. The color represents the treatment effect difference between the treatment/control arms. The color varies from black to yellow (red in the middle) where the variation ranges from -3 to 3.

This graphical approach satisfies C1, C2, C5 and partially C3 but fails to hold C4. Such LP only displays pairwise overlay of marginal subgroups rather than all overlap across subgroups. In addition, although unusual effect sizes across subgroups can be shown (if a large colour difference exists), it is unable to detect heterogeneity. The reason is that neither the overall effect size nor the estimation variation of effect size for subgroups are displayed. It is noted that only two covariates can be considered in a LP. Though the number of the marginal subgroups of each covariate can be easily ten (therefore, the subgroup number can reach to a hundred), this may lead to a subgroup sample size being zero. Moreover, because the cut-off points for continuous covariate are arbitrary, LP is more suitable for categorical covariates.

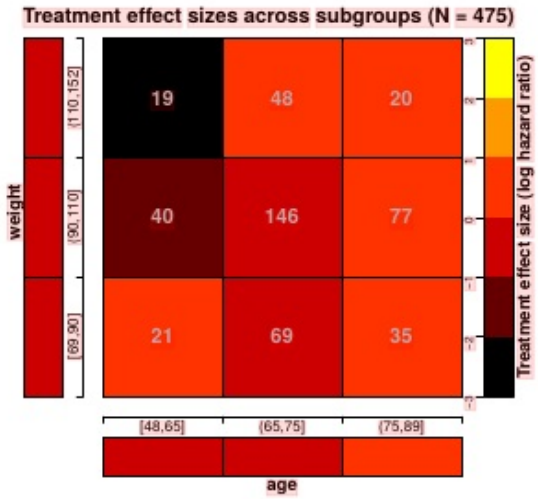


Figure 1: The level plot of treatment effect differences across 9 mutually disjoint subgroups defined by covariates age and weight. The cells on the bottom and the left margins are the marginal subgroups corresponding to the levels of age and weight. The inner figures of the cells stand for the subgroup sample sizes.

3.2 Contour plot

Figure 2 shows treatment effect differences in a particular region of age and weight by contour plots (CP). The contour lines are drawn through a bivariate interpolation and smooth surface fitting for irregularly distributed data points at prespecified grid points. Contour lines have an attached number showing treatment effect difference. The points inform that the corresponding subgroups have genuine effect sizes, where the subgroups are from the part of dataset and meet a certain condition on the values of covariate. The colours indicate the effect sizes of subgroups. Subgroup sample sizes and the overlap are only annotated in the bottom of the figure.

CP matches 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 neighboring 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. Also, it has no function of detecting heterogeneity because the overall effect size is not given.

There are few more noticeable characteristics for this graphical technique. CP is particularly useful when a dataset size is rather large and for variables 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. It is unclear how smooth the interpolated surface should be when the dataset size is not large but the values of two covariates are roughly distributed over the region

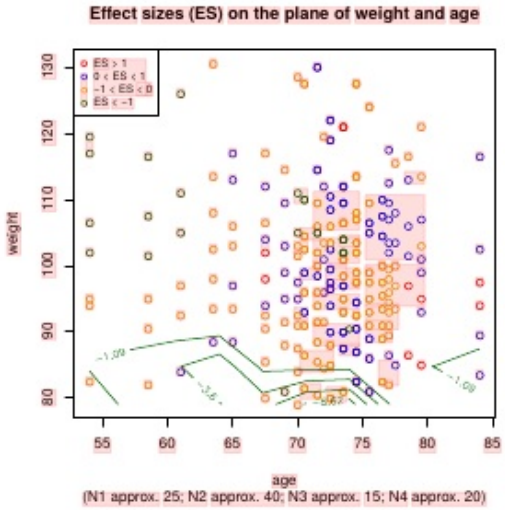


Figure 2: The contour plot of treatment effect differences over the plane of age and weight. N11 stands for the sample size of a marginal subgroup defined by a range of age; N12 means the overlap size of the immediate marginal subgroups on age; N21 is the sample size of the subset of a marginal subgroup on age but further defined by a range of weight; N22 represents the overlap size of the immediate subgroups (which are the subset of a marginal subgroup on age) on weight.

3.3 Venn Diagram

The Venn diagram (VD) for three subgroups defined by bone metastasis (bm), history of cardiovascular events (hx) and performance (pf) is shown in Figure 3. This diagram indicates the sample sizes for all the subsets forming by set operations (intersection and complement) on the three subgroups. The number outside of the three circle indicates the union of the three subgroups has a complement with size of 211.

VD apparently satisfies C2, C3 but not C1 and C4. Since information about treatment effect differences in subgroups is unavailable, VD do not allow to detect treatment effect heterogeneity.

VD also holds C5 because one can make a VD for any number of sets based on the Edwards construction [7, 8]. The total number of subgroups including mutual disjoint ones can be 2^n, where n is the number of the sets considered. Despite this merit there is be a limit on the number of the sets considered in practice. It may become more complicated to interpret a VD with more than five sets.

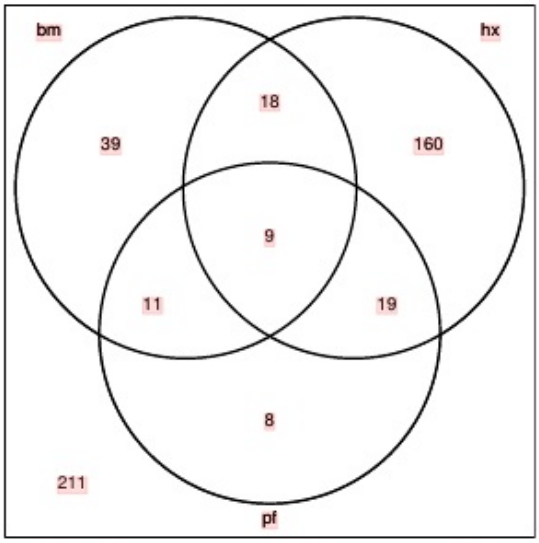


Figure 3: The Venn diagram of 3 subgroups defined by presence bone metastasis (bm), history of cardiovascular events (hx) and performance status = 1 (pf).

3.4 Bar Chart

A bar chart (BC) for treatment effect differences in subgroups defined by age and weight is shown in Figure 4. 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 bar stand for the treatment effect differences between the treatment/control arms, while the width is proportional to the square root of a subgroup sample size over the total square root sum. The colour merely shows which subgroup has the same category level on age. The length of the error bar above each colour bar exhibits the magnitude of the standard error of the point estimator.

This graphical representation approach holds C1, C2 and C5, partially C3 but not C4. Like LP, each bar is also the pairwise overlap of two subgroups defined by age and weight with their respective levels. Therefore, BC only provides 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, BC can handle more than ten (mutually disjoint) subgroups through increasing the level number of each covariate.

Few noteworthy characteristics also need to be mentioned. First, BC exhibits the standard errors of the point estimator for subgroup effect sizes. Second, it only considers two covariates. If considering few more covariates, one could label all the covariate's level combinations in the bottom part of the picture or simply to make a legend elsewhere. Third, it satisfies C5 though a high number of covariates or levels may cause a visualization problem (difficult to see how wide of the bar is). Fourth, it has the same issue of LP about the cutoff points for continuous covariates and is therefore favoured for categorical covariates.

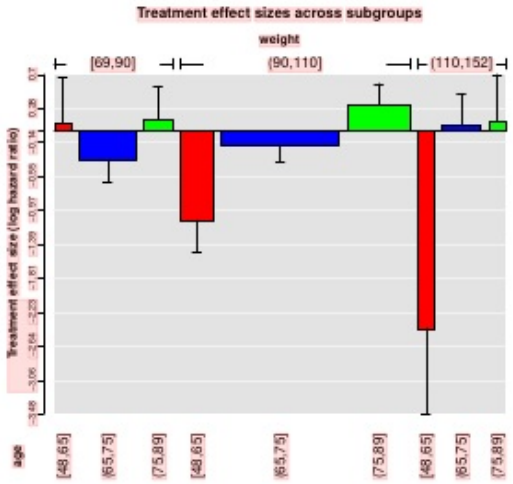


Figure 4: The bar chart of 9 mutually disjoint subgroups defined by the levels of age and weight

3.5 Forest Plots

A forest plot (FP) is a common graphical display approaches for meta analysis and subgroup analysis. Figure 5 shows one of its applications to the estimation of treatment effect differences in various subgroups. Here subgroups and their complements are defined by four binary covariates. The text on the left side 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 lines represents 95% confidence intervals of effects sizes (for subgroups) or treatment effects (for treatment/control arms). The square size reflects how a subgroup sample size is proportional to the full population. The solid vertical line for examining heterogeneity is located at the overall effect size as suggested in [9]. If there is a C.I. of subgroup effect not crossed by that solid line, we regard heterogeneity may occur in such a circumstance.

From the above description FP apparently holds all the criteria but C3 because of the inability to show subgroup overlaps. Note that the merit of showing treatment effect estimate for the treatment/control arms benefits practitioners in certain circumstance. Particularly, it is critical to prevent the subgroup that both interventions have harmful effects despite the promising effect size.

Note that the visual judgment on heterogeneity is slightly different from those in [1, 9, 10]. We later adopt the same recognition rule for the graphical approaches with similar design features.

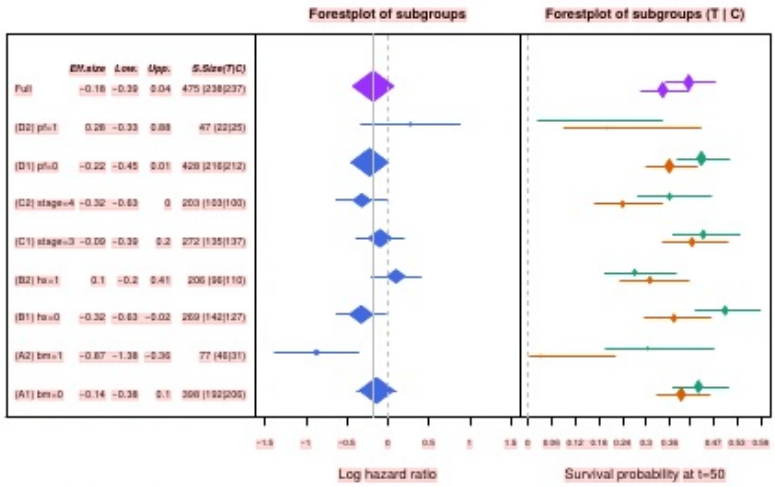


Figure 5: The forest plot for subgroups defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm). Effect sizes in terms of the log-hazard ratio and associated treatment and control group survival probabilities at time=50 are displayed.

3.6 Tree Plot

Figure 6 shows a tree plot (TP) of treatment effect differences for subgroups defined by bm, pf and hx. Each covariate has two categories, either the value is 0 or 1 (denoted by blue and red lines respectively). The tree is expanded from the full population to the subgroups defined by bm's categories. The subgroups are further divided into the subgroups defined by all the category combinations of hx and pf. If more variables were included, this division procedure is consecutively conducted to form subgroups until all the category combinations of the covariates are considered. Each layer shows the 95% confidence intervals (C.I.) of treatment effect differences for the associated subgroups where the scales are different layer by layer. 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.

TP matches all the criteria. It is obviously fit to C1, C2 due to the graphical designs. In addition, the subgroups at the same layer are formed by all possible set operations (intersection and complement) on the categories of the associate covariates. This feature is similar to VD and TP 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 FP, the assessment demands drawing an auxiliary horizonal 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, TP can certainly address more than 10 subgroups. Note that the number of subgroups also dependeds on how many covariates are involved and ategories each covariate has. A few features of TP are worthily pointed out. First, it provides information of the interval estimation for subgroup effect sizes. Second, it is possible to consider few more categories for each covariate by adjusting TP for visualizing all the effect sizes for associated subgroups. But, ideally and relatively this may need to reduce the number of covariates. Third, the maximum number of the covariates considered could be up to 5, otherwise a visualization problem may emerge. Fourth, TP have the same issue about cut-off points for a continuous covariate as LP and BC.

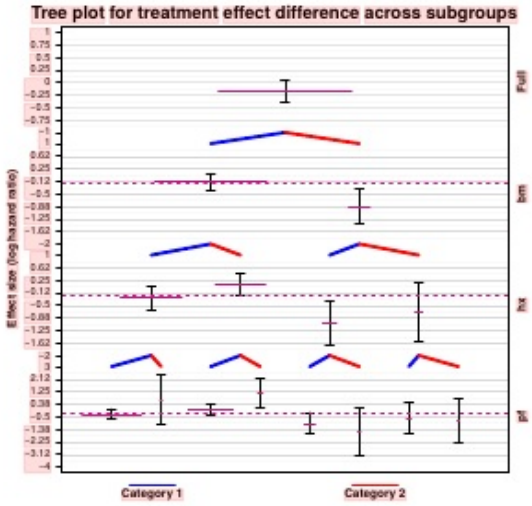


Figure 6: The tree plot of treatment effect difference for subgroups defined by all category combinations of the covariates existence of bone metastasis (bm), history of cardiovascular events (hx), and performance (pf). Each layer shows the 95% C.I. of treatment effect differences for the associated 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.

3.7 Galbraith plot

A Galbraith plot (GP) [11, 12] is an alternative or supplementary to a forest plot for examining heterogeneity of studies or subgroups in meta analysis. Its variant shown in Figure 7 exhibits the estimation of treatment effect sizes for subgroups defined by three covariates. The horizontal axis represents inverse standard error (1/SE), and the vertical axis stands for standardized estimates (namely a subgroup's effect size divided by the corresponding standard error). The gray band serves to examine heterogeneity if one standardized estimate is located outside the band. The central line points to the standardised estimate of the average effect size for the full population. Moreover, the arc is for effect sizes. A subgroup's effect size is registered at the red icon projected on the arc by the line from the origin through the corresponding point.

The result of GP's graphical assessment is satisfactory. Obviously, it holds C1, C4 and C5 because of its design features. It can handle much more subgroups and is also helpful to detect an outlier. GP only partially fit to the criterion C2. It only indirectly reveals information of subgroup sample sizes through individual standard errors. Moreover, it does not hold C3. Like FP, it is not possible to know subgroup overlap information.

The GP approach here has one difference in the central line location from the initial design. The line location is originally set at the average estimate of subgroup effects, where the pooled average with the proportion $\frac{1/SEi}{\sum_i 1/SEi}$ as a weight for Subgroup i. The average estimate is obtained by fixed effect models, where the truth effect sizes of all subgroups are assumed.

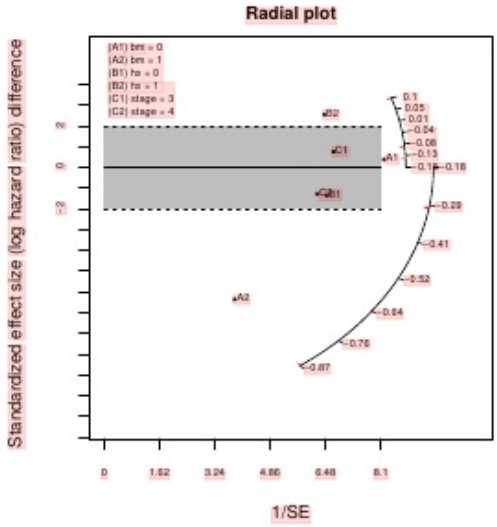


Figure 7: A Galbraith plot across subgroups and their complements defined by stage, history of cardiovascular events (hx) and existence of bone metastasis (bm)

3.8 L'Abbé plot

A L'Abbé plot (LAP)[13] is a variant of scatter plots which is feasible and useful for examining heterogeneity in meta analysis. The graphical design is originally for binary outcome data to represent a risk ratio, or a risk difference or an odds ratio between treatment and control arms. Here, we extend this graphical technique to the case of continuous and survival outcomes and also modify points to rectangles in Figure 8.

As seen, each subgroup has one estimate located at a position where its x-coordinate and y-coordinate correspond to the estimates of the control/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. Each rectangle has a vertical segment from its center to the diagonal dashed line which represents no effect size within a subgroup. The colour of a segment signals whether the corresponding treatment effect difference is positive (blue) or negative (red) or not. All the mean estimates of subgroup effect sizes are written in the left-top and right-bottom corners of the picture. Furthermore, the solid line parallel to the diagonal line has a y-intercept at the overall effect sizes.

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 FP.

AP shares the same graphical assessment results with FP. It satisfies all criteria except C3 for not showing subgroup overlap information. Two characteristics should be noted. First, it may handle as many subgroups as FP does. But, it may be difficult to recognize subgroups and its corresponding rectangles if more subgroups have close effect estimates 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

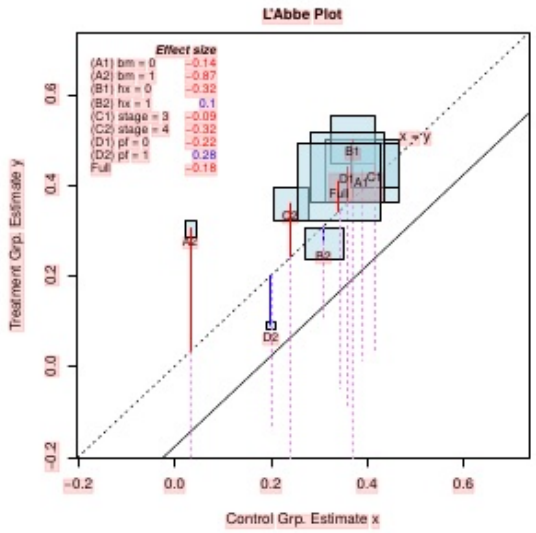


Figure 8: A L'Abbé plot for the subgroups and their complements defined by performance (pf), stage, history of cardiovascular events (hx) and existence of bone metastasis (bm)

3.9 STEPP approaches

The subpopulation treatment effect pattern plot (STEPP) method[14, 15] is of some publicity in breast cancer recently. It is a non-parametric method mainly for examining whether treatment-covariate interactions exist.

In Figure 9, we adopted the slide-window fashion of STEPP to represent the estimation of treatment effect differences in overlapping subgroups defined by the covariate age. Each subgroup has a sample size of around 100 and also has about 80% being overlapped with the neighboring 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.. The red line is formed by connecting the mean point estimates of treatment effect difference for all individual subgroups. The green represent the mean point estimate of treatment effect difference 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 may reveal interaction exists.

STEPP has a reasonable graphical assessment result –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 approach has two flaws. One is a strong restriction on considering one continuous covariate. It is difficult to extend the application for more continuous covariates. Another is 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 range of the sample sizes for subgroup and overlap. The analysis results are further compared with the graphical results by using MFPI algorithm [16, 17] or non-parametric methods (such as Gaussion processes [18]), where a functional curve of the covariate on treatment effect is interpolated.

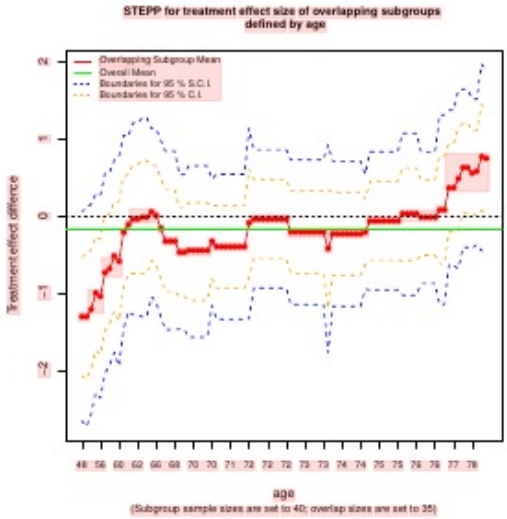


Figure 9: The STEPP plot of overlapping subgroups defined by age. Each subgroup has a sample size of around 100 (N11 = 100) and is controlled to have about 80% (N12/N11) being overlapped with the neighboring subgroups.

3.10 Alluvial diagram

Alluvial diagrams are flow diagrams that can be used to display the distribution of the subjects across the subgroup defining covariates. Figure 10 shows the alluvial plot for the subjects in the prostate cancer dataset. The red coloured bands correspond to patients that were randomized to treatment (rx = 1) while blue bands to patients in control (rx = 0). 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).

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.

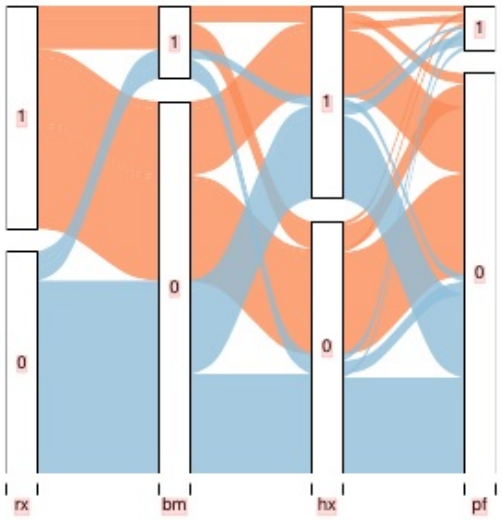


Figure 10: Alluvial plot displaying the distribution of patients across the subgroups. The red coloured bands correspond to patients that were randomized to treatment (rx = 1) while blue bands to patients in control (rx = 0). The width of the bands are proportional to the sizes of the subgroups.

3.11 UpSet Plot

UpSet plots are a novel visualization technique for the quantitative analysis of sets and their intersections [19]. In Figure 11, we used the `UpSetR` package [20] 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 the advantage of the UpSet plots 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 4.1. 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.

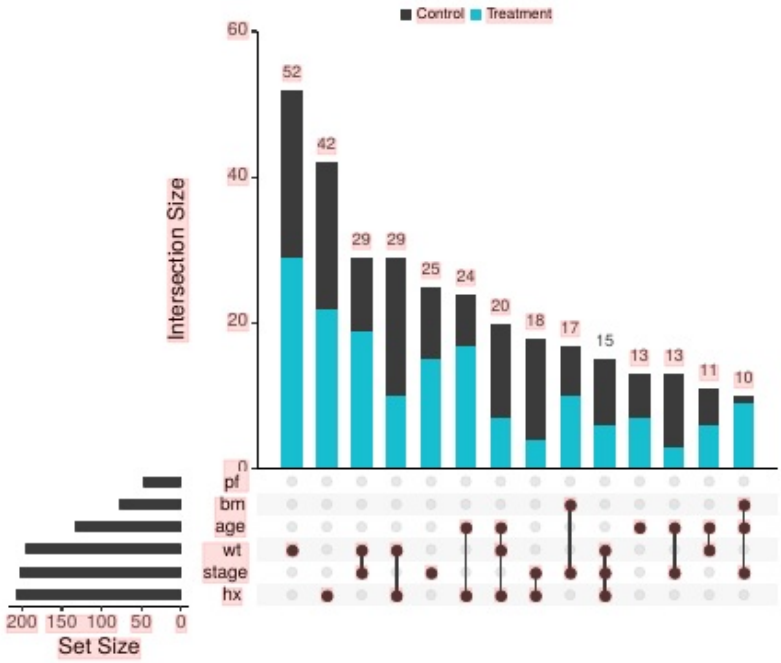


Figure 11: Upset plot displaying the subgroups conformed by the instersection of all subgroup defining covariates and their sizes.

4 Improvement and alternatives

The merits and demerits of the nine methods were assessed previously based on the criteria. In this section we focuses on improving several potential improvement on level plots, Venn diagrams, forest plots, Galbraith plots, L'Abbé plots and UpSet plots.

4.1 Improved UpSet plot

A possible improvement to the UpSet plot is displayed in Figure 12. We extend the UpSet package to display effect sizes in an extra panel. In this case, as we are working with survival data, the log hazard ratio and its confidence interval is shown. The overall treatment effect and its confidence interval is also included in red. This information is similar as that on the forest plots. However, the UpSet plots provides the advantage to observe intersection of sets, arrange them in terms of their sizes, and display the treatment effect.

This extension of UpSet plot also allows to display lower level intersections. The matrix panel have three levels with a new icon: a '+' symbol if variable is equal to 1, '-' if variable is equal to 0, 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 considers 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.

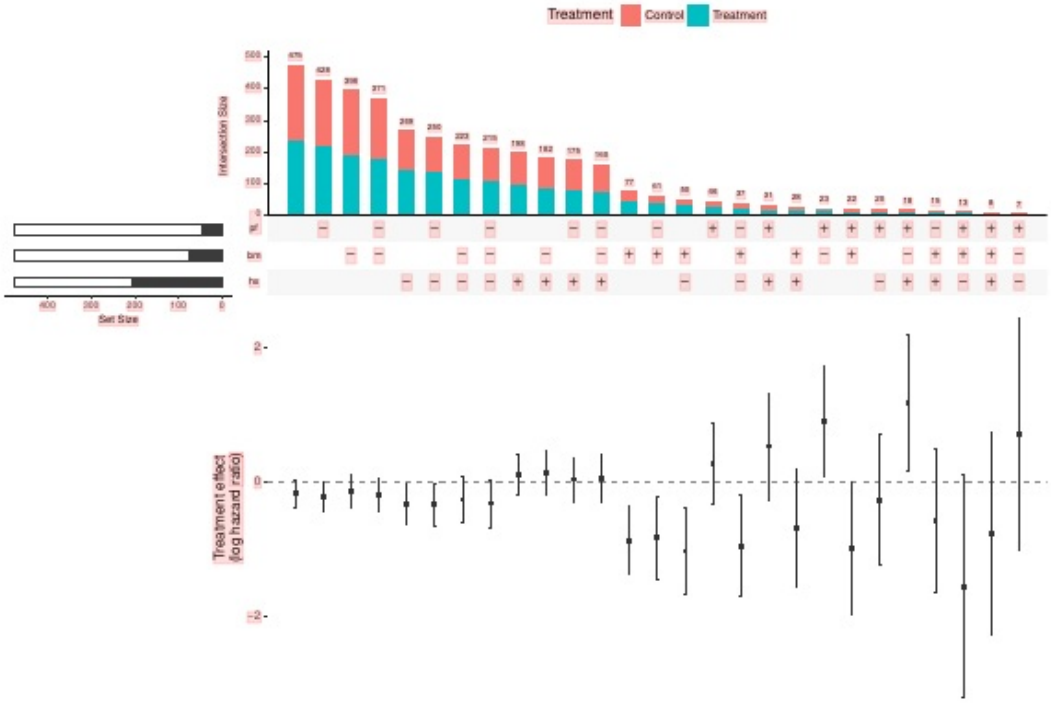


Figure 12: Improved UpSet plot for subgroups defined by performance (pf), bone metastasis (bm) and history of cardiovascular events (hx). Subgroups sizes are displayed on the panel on top with the bar plot, while the treatment effect sizes in the panel on the bottom. The panel in the middle displays how the subgroups are conformed by assigning a '+' if the variable is equal to 1 and a '-' if the variable is equal to 0

4.2 Improved Level plot

Figure 13 shows the improved LP which inherits all merits and improves the second demerit of the previous LP (Figure 1). The size of the coloured square inside each cell represents the proportion of the subgroup sample size to the full population. This new design feature allows one to recognise the subgroup sample sizes more easily. But, it may be difficult to see what colour is in each square, particularly in the case of small sample sizes.

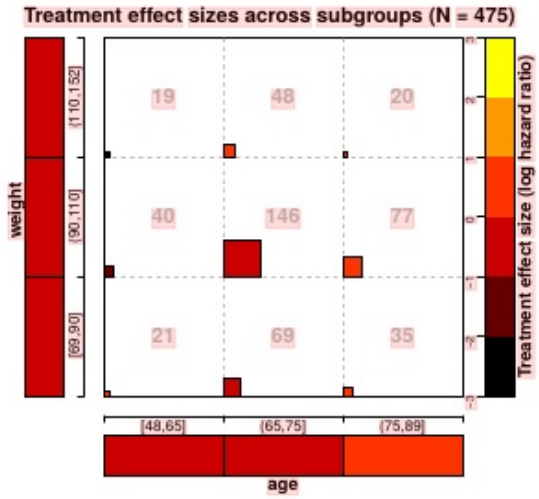


Figure 13: The improved levelplot of treatment effect difference across all 9 mutually disjoint subgroups defined by Age and Weight

4.3 Improved Venn Diagrams

Figure 14 and 15 are the improved VD considering four and three subgroups, respectively. Both represent the treatment effect differences of subgroups by colouring the corresponding regions, where the magnitude is shown in the right colour bar. This feature thus enables improved VD 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 14, using four ellipses for representing all possible subgroups (formed through intersection and complement) is visually appropriate. Other patterns (such as polygons [21, 22]) can be also applied but the visualisations may not be easily understood, comparable with that shown in Figure 14.

Figure 15 further considers area-proportional methods, where each subgroup's representative region area is proportional to the respective sample size proportion. Here we employed the simple algorithm mentioned in[23]. The region areas in this instance only approximately correspond the sample size proportions. This is possibly because of the limited degrees of freedom for circles.

In fact, accurate displays of the region areas for the sample size proportions is achievable. Recently Micallef and Rodgers developed an algorithm which can produce an accurate area-proportional Venn diagrams using ellipses [23]. However, their algorithm is somewhat sophisticated and only can work on three sets.

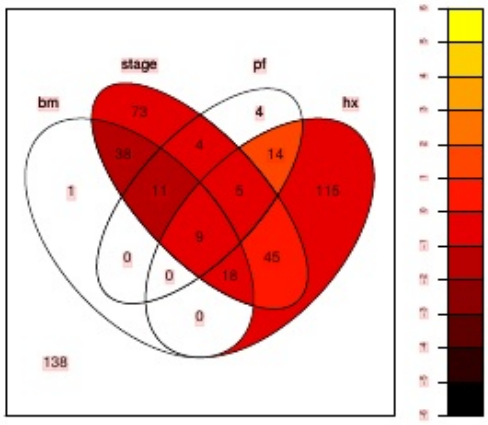


Figure 14: The improved Venn diagram of 4 sets defined by presence of bone metastasis (bm), disease stage, performance status = 1 (pf) and history of cardiovascular events (hx)

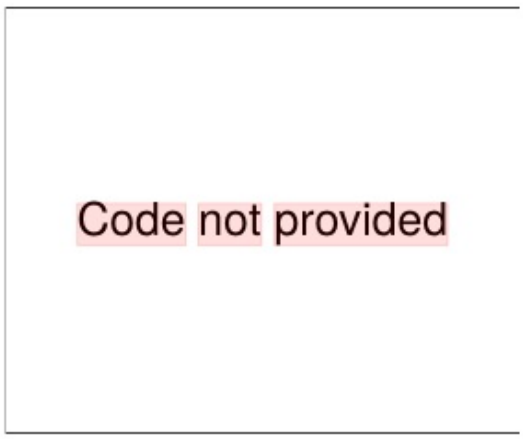


Figure 15: The approximate area-proportional Venn diagram of 3 sets defined by specified ranges of V1, V2 and V6.

4.4 Alternatives

Forest plots, Galbraith plots and L'Abbé plots share the same failing on incapacity of showing subgroup overlaps. One potential improvement measure is to consider combing relevant figures about overlap information.

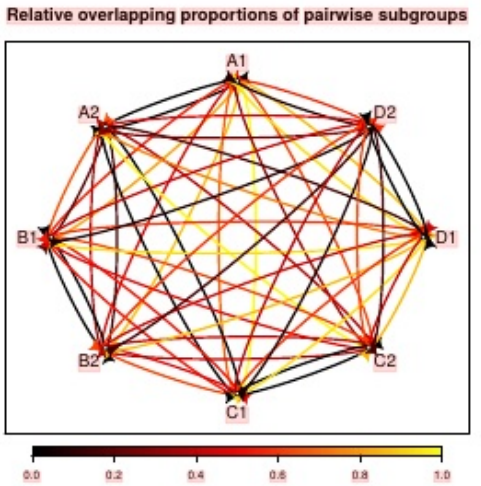
The plots shown in Figure 16 exhibit certain subgroup information about pairwise overlap proportions or similarity measures. Figures 13a-13d show pairwise relative overlap proportions, where different colours show the range of overlap magnitude.

More specifically, Figure 16a 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 16b is a variant of Figure 16a. Two identical sets of subgroup labels around two circles and each shows relative overlapping proportions with unidirectional arrowed coloured lines. Unlike Figure 13a, the arrow position is now near the end. The subgroup labelled at the starting point of the arrowed line is a baseline subgroup for the relative overlapping proportion. Figure 16c 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 16d 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. There are two variants of Figure 16b and 16d shown in appendix.

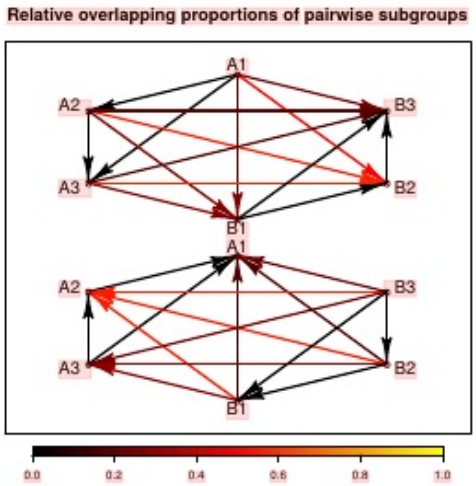
Both Figures 16e-16f show dissimilarity distance, which is defined by one minus a relative overlap proportion. Each line of Figure 16e shows the dissimilarity distance of a subgroup with the others, where the baseline subgroup is denoted by a green triangle. The red crosses below a line are located according to actual dissimilarity distances; the red subgroup labels above each line segment correspond to the red crosses, where the labels are placed by order based on their actual dissimilarity distances. Figure 16f shows the same information as Figure 16e, where the coloured lines represent subgroups. There are one variant of Figure 16e shown in appendix. Note that for each subgroup we do not show its dissimilarity distance to itself and its complement.

For our improvement task, we present an example of the combined forest plot and matrix plot for ten subgroups and their respective complements in Figure 17. As seen, the visualisation is appropriate and the outcome displays subgroup overlap information. Similarly, an improved version of GP and AP can be created by attaching any plots in Figure 16e. While the hybrid graphical representation improves the original demerits, it may lead to a recognition issue upon the large number of subgroups. Another alternative to improve FP, GP and LAP is simply to combine a VD. This measure can provide full overlap information of all subgroups. But, regarding effective visualisation one may only consider a small number of subgroups (up to five) for the hybrid graphical representation.

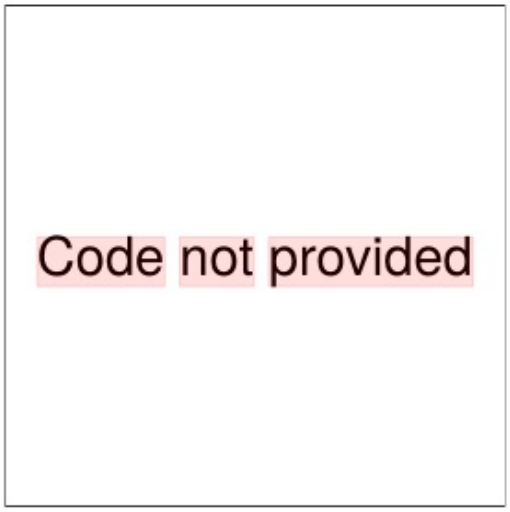
Incidentally, the Jaccard index, namely $|A \cap B|/|A \cup B|$ for any sets A, B, can replace pairwise overlap proportions for subgroup overlap information. The graphical display is thus simplified due to not showing repetitive Jaccard indexes. However, this measure may lead to missing some information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



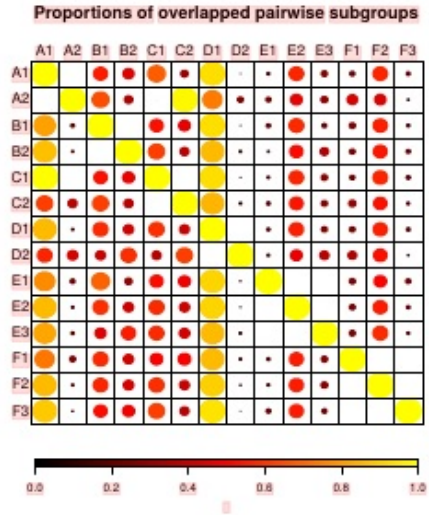
(a) Line plots with bidirectional arrowed curves for relative overlap proportions for pairwise subgroups.



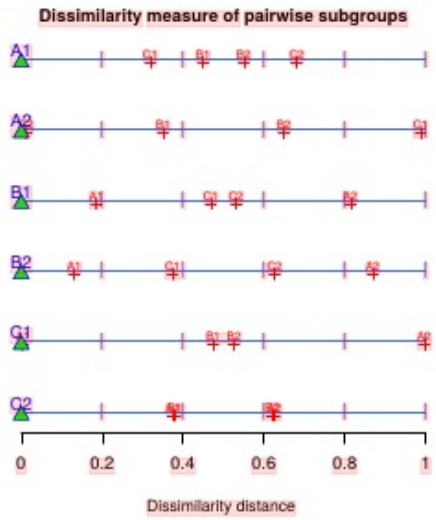
(b) Line plots with unidirectional arrowed lines for relative overlap proportions for pairwise subgroups.



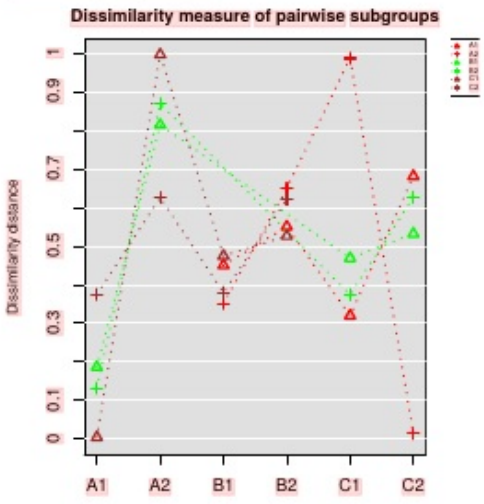
(c) Line plots for relative overlap proportions for pairwise subgroups.



(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 16: Plots for subgroup information about pairwise overlap proportions or dissimilarity measure.

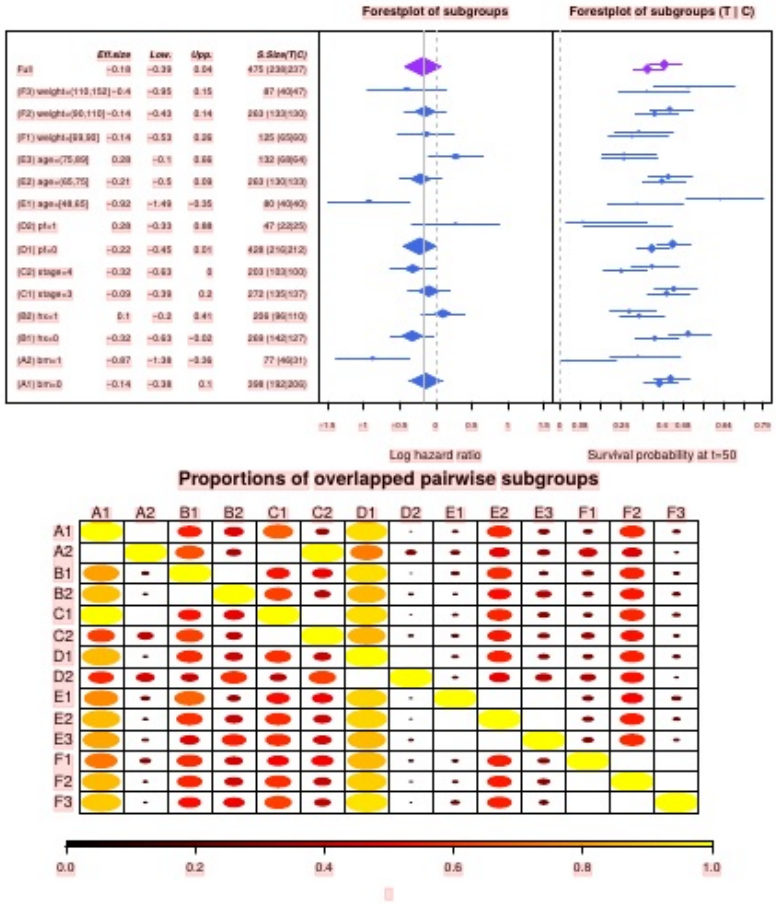


Figure 17: An improved forest plot across subgroups, the complements and the associated treatment/control groups. The subgroups are defined by performance (pf), stage, history of cardiovascular events (hx), existence of bone metastasis (bm), age and weight

Table 1: The assessment summary of graphical techniques for subgroup problems. The assessment criteria are: 1. whether to display effect sizes for subgroups (**C1**); 2. whether to show subgroup sample sizes (**C2**); 3. whether to exhibit all subgroup overlap information (**C3**); 4. whether to serve for detecting heterogeneity in subgroup effect sizes (or the treatment-covariate interaction) (**C4**); 5. whether is available for the large number of subgroups (more than 10) (**C5**). The subscript * of some graphical approaches denote they have been improved. The exclamation mark denotes notable points.

Criterion	LP*	CP	VD*	BC	TP	FP*	GP*	LAP*	STEPP	Alluvial	UpSet*
C1	✓	✓	✓*	✓	✓	✓	✓	✓	✓		✓
C2	✓*		✓	✓	✓	✓	Δ	✓		✓	✓
C3	Δ		✓	Δ	✓	Δ*(!)	Δ*(!)	Δ*(!)		✓	✓
C4					✓	✓	✓	✓	✓		✓
C5	✓	✓	✓(!)	✓	✓(!)	✓	✓	✓	✓	✓	✓

Table 2: The feature summary of graphical techniques for subgroup problems. The abbreviation (Info. for C.I.) means whether there is available information to build a confidence interval; T/C effect means the approach presents the treatment effects of treatment/control arms; N_c means the number of covariates for considerations; P.O./A.O. stands for pairwise overlay or all overlap for subgroups. The subscript * of some graphical approaches denote they have been improved. The exclamation mark denotes notable points.

Feature	LP*	CP	VD*	BC	TP	FP*	GP*	LAP*	STEPP	Alluvial	UpSet*
Info. for C.I.				✓	✓	✓	✓	✓	✓		✓
T/C effect						✓		✓			
P.O./A.O.	P.O.	P.O.	A.O.	P.O.	A.O.	P.O.(!)	P.O.(!)	P.O.(!)	P.O.	P.O.	A.O.
N_c	2	2	2-5	2-5	1-5	1-40	1-1000	1-40	1	2-10	2-100

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 and 2.

The general summary is as follows: all the nine graphical techniques satisfy the primary criterion about displaying subgroup effect sizes. Except LP, CP and VD, the rest displays or has information to construct confidence intervals, specifically BC and GP exhibit standard errors for estimators. Furthermore, only two (FP and LAP) 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 GP indirectly show the information through the standard error of estimators.

The third criterion is fully and partially hold for all apart from CP and STEPP. VD and TP show the overlay of all subgroups. But, the remaining approaches only display the overlay for pairwise subgroups. Six graphical displays featuring different types of lines and design characteristics were invented for improving FP, GP and LAP by showing pairwise subgroup overlay. It is noted that when the number of subgroup is small (say, up to five), the improved FP, GP and LAP can combine a VD 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, VD and TP practically can deal with only up to five sets (considered for overlap) for effective visualisation. Even an area-proportional VD can afford merely three sets. Moreover, six approaches are able to regard a small number of covariates for subgroups. Only FP, GP and LAP 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, CP can be particularly useful when a data set is large and the distributions of two covariates considered are roughly uniform; LP and BC may be easier for some audience to understand subgroup information due to their simple design; FP and LAP which show the

treatment effects of two comparative intervention arms 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 [24] in epidemiology before utilising the graphical approaches.

In addition, the focus on developing a two-dimensional graphical display can be contentious. We virtually undoubt 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.

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

[8] 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.

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

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

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] 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.

[22] 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.

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

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