A framework to assess the value of subgroup analyses when the overall treatment effect is significant

Hong Sun and Werner Vach

Clinical Epidemiology, Institute for Medical Biometry and Statistics, Medical Center-University of Freiburg, Freiburg, Germany

ABSTRACT
Although subgroup analysis has been developed and widely used for many years, it is still not clear whether we should perform and how to perform such subgroup analyses when the overall treatment effect is significant. In this paper, we develop a framework to assess and compute the long-term impact of different strategies to perform subgroup analysis. We propose two performance measures: the average gain for patients in the future (E) and the probability of recommending a change to a worse treatment at individual patient level (P). Five families of decision rules are applied under different assumptions for the individual treatment effect (TE) variation. Three distributions reflecting optimistic, moderate, and pessimistic scenarios are assumed for true treatment effects across studies. This framework allows us to compare subgroup analyses decision rules, and we demonstrate through simulation studies that there are decision rules for subgroup analysis which can decrease P and increase E simultaneously compared to the situation of no subgroup analysis. These rules are much more liberal than the usual superiority testing. The latter typically implies a dramatic decrease in E.

1. Introduction
The choice of relevant patient populations is an important step in the benefit assessment of a new drug. It may result in the application of subgroup analyses in a clinical trial, even if the overall effect was significant. A nice illustration of this point is the debate regarding the effect of Clopidogrel in the CAPRIE trial (Bender et al., 2010; Hasford et al., 2010). The trial was a randomized triple-blind clinical trial including patients with atherothrombosis, i.e., diagnosed by the disease manifestations of myocardial infarction (MI), stroke, or symptomatic peripheral arterial disease (PAD), randomly assigned to be treated by either Clopidogrel or Aspirin using block randomization with center and disease group (Stroke, MI, and PAD) as stratification factors. The main objective was to test the superiority of Clopidogrel vs. Aspirin in the secondary prevention of vascular events. The primary endpoint was the first occurrence of an event either MI, ischemic stroke, or vascular death. The primary analysis defined in the protocol was an intention-to-treat (ITT) analysis for the primary endpoint in the total cohort of patients. A sample size of 19,185 was included to detect an overall relative risk reduction (RRR) of 11.6% with 90% power using a two-sided log-rank test with 5% significance level (CAPRIE Steering Committee, 1996).

The results from the ITT analysis showed a statistically significant \( p = 0.043 \) TE in terms of RRR of 8.7% in favor of Clopidogrel (95% confidence interval [CI] 0.3–16.5), which was consistent
with the assumed TE (11.6% RRR) in sample size calculation. In an additional analysis, the CAPRIE investigators separately examined the effect of the treatments on the primary outcome in each of the three strata of disease group (Fig. 1), suggesting some heterogeneity. Formally, this corresponded to a borderline statistically significant interaction \( p = 0.042 \) (CAPRIE Steering Committee, 1996).

Interestingly, based on these two analyses, several institutions came to different decisions. In the approval of Clopidogrel for the European market, European Medicines Agency (EMA) accepted the primary endpoint analysis of the overall population because neither strong heterogeneity nor a deficient definition for the overall study population has been found in the CAPRIE trial (European Medicines Agency, 2000; Hasford et al., 2010). Similarly, FDA adhered to the ITT analysis for CAPRIE and approved Clopidogrel accordingly for all patients with atherothrombotic diseases (Food and Drug Administration, 1998). In a cost–benefit assessment for the health care system in the UK, NICE concluded with respect to efficacy in accordance with the primary ITT analysis of the overall population of the CAPRIE trial. However, it considered the balance between clinical effectiveness and cost-effectiveness not to justify a replacement of aspirin by Clopidogrel to prevent vascular events (NICE, 2005). In a benefit assessment for the German system, IQWiG acknowledged the superiority of Clopidogrel only in the subgroup of patients with PAD but performed no evaluation of cost-effectiveness (IQWiG, 2006).

The desire to perform subgroup analysis after reaching a significant overall effect appears also in multi-regional clinical trials (MRCT), where the overall effect is the global TE in the whole population, regardless of the regions, but since the same treatment may work differently in patients from different countries and regions, subgroup analysis by country or region is often required by regulatory authorities in MRCTs (Pocock, 2013). However, the situation is slightly different here, as the subgroup analyses are already intended when planning the study, and can be taken into account in sample size calculation (Ikeda and Bretz, 2010).

Although subgroup analyses in clinical trials have been a matter of debate and methodological investigation for at least two decades (ICH Steering Committee, 1999; Pocock et al., 2002; Yusuf et al., 1991), we now meet a rather new and different situation. Traditionally, subgroup analyses were discussed in the context of trials failing to demonstrate an overall effect, and the focus was on limiting the risk of generating spurious signals of treatment superiority by testing many (small) groups and focusing too much on the estimates of maximally observed effects. If we apply subgroup analyses after

![Figure 1. RRR and 95% CI by disease subgroups in CAPRIE trial. The figure is reproduced from CAPRIE Steering Committee, 1996 with kind permission of "The Lancet."](image)
demonstrating an overall effect, we may be more concerned about failing to demonstrate an existing TE in a subgroup due to insufficient power. There may be too much focus on confidence limits or the observed minimal effects, even if the spread of the treatment effect over subgroups may mainly reflect a random fluctuation. This new situation did not yet attract much attention, although an EMA guideline proposal on subgroup analysis in 2010 already mentioned the topic and the 2014 version of the draft guideline includes a section on this (European Medicines Agency, 2010, 2014).

The main objective of this paper is to provide and use a framework to assess and compute the long-term effect of different strategies for performing subgroup analyses when the overall TE is significant. The formal framework and decision rules considered are presented in Section 2. Results are displayed in Section 3. Section 4 summarizes main results and discusses potential limitations. A conclusion in Section 5 finalizes the paper.

2. Methods

2.1. Notations

Our framework is based on considering a sequence of clinical trials, all comparing some standard treatment to a new experimental treatment. The sample size calculation for all studies is based on the same assumed effect \( \theta_A \), resulting in identical sample sizes \( N \). The patient population of each study can be divided into \( K \) subgroups of equal size, and we have a decision rule \( \psi \) to decide on the superiority of the experimental treatment in the whole study population, and a decision rule \( \phi \) to perform such a decision at the subgroup level.

In the sequel, we denote the studies by \( s = 1, \ldots, S \), the subgroups in each study by \( g = 1, \ldots, K \), and the individual patients in each subgroup by \( i = 1, \ldots, I \) with \( I = N/K \). For each patient, we assume an individual true TE \( \theta_{sgi} \), because it is unrealistic for every patient to have the same characteristics apart from group partition, thus each individual patient would react differently to the same treatment. Further, \( \theta_{sg} \) denotes the true TE in subgroup \( g \) of study \( s \), and \( \theta_s \) denotes the true TE in study \( s \).

2.2. Performance measures

In evaluating the performance of particular subgroup decision rules, we propose two measures. The point of departure of these measures is the general goal of clinical trials to improve the overall outcome of patients by accepting new effective treatments for whole patient groups without putting too many patients at risk of being harmed due to heterogeneity of treatment effects within such groups. However, due to the stochastic nature of the results of clinical trials, we have to balance between these two subgoals. On one hand, we would like to improve the overall outcome as much as possible by accepting as many effective treatments as possible. On the other hand, we want to minimize or control the risk of recommending a patient to switch to a treatment, which is actually inferior, i.e., with \( \theta_{sgi} < 0 \).

To measure the overall improvement, we consider the overall gain. We assume that, if the experimental treatment is declared superior in a subgroup, all patients in that subgroup will receive the new treatment instead of the standard treatment in future. Hence, based on our simulation, this gain is defined as

\[
E = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{K} \sum_{g=1}^{K} \sum_{i=1}^{I} \theta_{sgi}^* \tag{1}
\]

with

\[
\theta_{sgi}^* = \begin{cases} 
\theta_{sgi} & \text{if } \phi_{sgi} = 1 \\
0 & \text{if } \phi_{sgi} = 0
\end{cases} \tag{2}
\]

where \( \phi_{sgi} \) denotes the result of applying the decision rule \( \phi \) in subgroup \( g \) of study \( s \).
In the case of $\theta$ being a difference in survival or success probability, $E$ approximates the average gain in survival or success probability over all future patients, if we always follow the recommendations made by the decision rules. If $\theta$ is a difference in the expected value of a continuous outcome, e.g., a quality of life score, then $E$ is the average gain in this score over all future patients.

To measure the risk of recommending an inferior treatment, we consider the fraction of patients with a negative true TE in the subgroups with a positive decision, i.e.,

$$P = \frac{\sum_{s=1}^{S} \sum_{g=1}^{K} \sum_{i=1}^{I} 1(\theta_{sg} < 0 \land \phi_{sg} = 1)}{\sum_{s=1}^{S} \sum_{g=1}^{K} \sum_{i=1}^{I} 1(\phi_{sg} = 1)}$$

(3)

$P$ approximates the probability of recommending an inferior treatment among all patients changing treatment as a consequence of the study results. In general, we aim to accept as many treatments with a positive effect as possible for as many patients as possible. At the same time, we should keep the risk of recommending new treatments to patients who do not benefit from the new treatment as low as possible. Consequently, we should aim at maximizing $E$ and minimizing $P$. Good decision rules should imply a reasonable balance between $P$ and $E$.

2.3. Subgroup decision rules

In this paper, we consider five families of subgroup decision rules. All these rules will be only applied if the overall rule $\hat{\phi}$ decides on superiority, i.e., if $\hat{\phi} = 1$.

In the first family $\phi^{<,\alpha}$ with $\alpha \in (0.0, 1.0)$, we consider the null hypothesis of usual superiority testing, namely that the new treatment is inferior to the standard treatment, i.e., $\phi^{<,\alpha}_{sg} = 1$ if and only if the lower bound of the two-sided $(1 - \alpha)$% CI for $\theta_{sg}$ is above 0.

In the second family $\phi^{\sim,\alpha}$ with $\alpha \in (0.0, 1.0)$, we take the opposite view: we do not require evidence for the superiority of the new treatment, but deny a subgroup effect only if we have evidence for the inferiority, i.e., $\phi^{\sim,\alpha}_{sg} = 0$ if and only if the upper bound of the $(1 - \alpha)$% CI for $\theta_{sg}$ is below 0.

For the limiting case $\alpha = 1.0$, in both families the decision rule reduces to comparing the estimate with zero, i.e., $\phi^{E}_{sg} = 1$ if $\hat{\theta}_{sg} \geq 0$. In the second family, for the limiting case $\alpha = 0.0$, we approach the situation that no subgroup analyses are performed, i.e., we decide on superiority for all subgroups as soon as we decide on overall superiority, i.e., $\phi^{N}_{sg} = \hat{\phi}$.

Since there is some tradition in subgroup analysis to only perform subgroup specific tests in the presence of evidence for heterogeneity of the subgroup specific TEs (Brookes et al., 2004), we also introduce the family $\phi^{l,\delta,\alpha}$. Here $\phi^{l,\delta,\alpha}_{sg} = 1$ if and only if the null hypothesis $H_{0}; \theta_{s1} = \cdots = \theta_{sn}$ can be rejected at level $\delta$ and if the lower bound of the $(1 - \alpha)$% CI for $\theta_{sg}$ is above 0.

Finally, we consider a family popular in the analysis of multiregional trials where the estimate in a subgroup should be at least some fraction of the overall TE estimate (JPFSB, 2007) $\phi^{F,\gamma}_{sg} = 1$ if and only if $\hat{\theta}_{sg} \geq \gamma \hat{\theta}_{s}$. Table 1 summarizes the main properties of the families considered. To well illustrate the difference among the five families of decision rules, we reconsider the CAPRIE trial, if applying the different decision rules to the CAPRIE trial, quite different decisions can be reached.

Starting from the first family $\phi^{<,\alpha}$ with $\alpha = 0.05$, we conduct subgroup analysis in the widespread used way to perform superiority testing in each subgroup at the 5% level, or equivalently to compare the lower bound of the 95% CI with 0. As we can see from Fig. 1, a treatment effect will be claimed only in the PAD group. If we weaken this criterion by allowing a larger $\alpha$ and consequently a narrower confidence interval, in the Stroke group, we observe a CI of $[-3.5, 17.0]$ for $\alpha = 0.1$ and of
Table 1. Summary of the families of subgroup decision rules.

| Family                  | Notation | Condition for \( \phi_{\alpha} = 1 \) |
|-------------------------|----------|-------------------------------------|
| 1 Superiority testing   | \( \phi^{<,a} \) | if and only if the lower bound of the two-sided \( 1 - \alpha \) CI for \( \theta_{sg} \) is above 0 |
| 2 Inferiority testing   | \( \phi^{>,a} \) | if and only if the upper bound of the \( 1 - \alpha \) CI for \( \theta_{sg} \) is not below 0. |
| 3.1 Comparing estimate  | \( \phi^{E} \) | if \( \theta_{sg} \geq 0 \) with 0 |
| 3.2 No subgroup analysis| \( \phi^{N} \) | if and only if the null hypothesis \( H_0: \theta_{n1} = \cdots = \theta_{ns} \) can be rejected at level \( \delta \) and if the lower bound of the \( 1 - \alpha \) CI for \( \theta_{sg} \) is above 0 |
| 4 Interaction gatekeeper| \( \phi^{I,\delta,a} \) | if and only if the null hypothesis \( H_0: \theta_{n1} = \cdots = \theta_{ns} \) can be rejected at level \( \delta \) and if the lower bound of the \( 1 - \alpha \) CI for \( \theta_{sg} \) is above 0 |
| 5 Fraction of estimate   | \( \phi^{F,\gamma} \) | if and only if \( \theta_{sg} \geq \gamma \theta_{s} \) |

[0.04 14.0] for \( \alpha = 0.26 \). Hence, we can now claim a treatment effect both in the PAD and the Stroke groups, if \( \alpha \geq 0.26 \). This also holds for the limiting case \( \alpha = 1.0 \), i.e., for \( \phi^{E} \), where we only compare the estimate with 0, as (only) in these two groups the effect estimates are above 0 (cf. Fig. 1).

Now we can even become more liberal and consider the second family \( \phi^{>,a} \), comparing the upper bound of the \( (1 - \alpha)\% \) CI with 0. We start with \( \alpha = 1.0 \), i.e., with \( \phi^{E} \), and then decrease \( \alpha \). In the MI group, for the \( \alpha = 0.68 \), the 32% CI is \([-7.3, -0.2]\), i.e., the upper bound of MI group starts to be negative. Hence for \( \alpha < 0.68 \), we would accept Clopidogrel in all three groups.

Since the interaction between treatment effect and disease group is statistically significant \((p = 0.042)\), using decision rule family \( \phi^{I,\delta,a} \) with \( \delta = 0.05 \), is equivalent to \( \phi^{<,a} \). Consequently for \( \alpha = 0.05 \), Clopidogrel is only accepted in the PAD group. When using the last family of decision rules \( \phi^{F,\gamma} \) with \( \gamma = 0.5 \) as suggested in the Japanese guideline of MRCTs (JPFSB, 2007), we would reject Clopidogrel only in the MI group, as not only the largest estimate found in the PAD group, but also the estimate of 7.3 observed in the Stroke group is larger than half of the overall estimate, i.e., \( 8.7/2 = 4.35 \) (CAPRIE Steering Committee, 1996).

2.4. Assumptions on subgroup specific TEs and individual TEs

Given the study specific TE \( \theta_s \), we assume for the subgroup specific effects

\[
\theta_{sg} \sim \mathcal{N}(\theta_s, \sigma^2_G) \tag{4}
\]

with \( \mathcal{N} \) denoting a centered normal distribution, ensuring that the average of the subgroup specific effects \( \bar{\theta}_s \) is identical to \( \theta_s \) (for more details see Appendix I of Supplementary Material). Within each subgroup, it would not be reasonable to assume identical TEs for each patient, because the subgroup categorization considered typically reflects only one factor such as age, gender, or a disease characteristic. Hence it is likely that there are also other factors with an influence on the treatment effect, resulting in an additional variation from patient to patient. Hence for the individual TEs we assume

\[
\theta_{sg} \sim \mathcal{N}(\theta_{sg}, \sigma^2_{GI}) \tag{5}
\]

again, ensuring that \( \bar{\theta}_{sg} = \theta_{sg} \).

To facilitate the interpretation of our results, we reparameterize the variance components \( \sigma^2_G \) and \( \sigma^2_{GI} \) in the following way: with \( \sigma^2_I = \sigma^2_{GI} + \sigma^2_G \), we denote the overall variance of the individual TEs in a single study (i.e., the conditional variance given \( \theta_s \)). We express the between-subgroup variation \( \sigma^2_G \) as a fraction of the overall variation of the individual TEs, i.e., we define

\[
R^2 = 1 - \frac{\sigma^2_{GI}}{\sigma^2_I} = \frac{\sigma^2_G}{\sigma^2_I} \tag{6}
\]

\( R^2 \) can be easily interpreted as an explained variation, i.e., how much of the inter-individual variation in TEs can be explained by between-subgroup variation. Further, we relate the overall
individual TE variation to the assumed effect by introducing \( \tau \), in order to obtain a quantity which is independent of the scale for the TE:

\[
\tau = \sigma_I/\theta_A
\]

The following interpretation of \( \tau \) may be helpful: If the true effect \( \theta_i \) is equal to the assumed effect \( \theta_A \) for a single trial, the choice \( \tau = 0.5 \) implies 2 times of standard deviation (SD) of \( \theta_{gi} \) equals \( \theta_A \), and hence corresponds to a situation where 2.5% of the patients have a negative TE. Similarly, \( \tau = 1.0 \) corresponds to a situation where about 15.8% of the patients have a negative TE.

### 2.5. Assumptions on distribution of true study effects

The success rates of clinical trials (in the sense of reaching a significant positive TE for the new treatment) varies substantially among different areas (Dent and Raftery, 2011; Djulbegovic et al., 2013, 2008; Kumar, 2005). Hence it is necessary to consider different scenarios for the true TE. We consider three scenarios in this paper.

In the first one, we assume that on average the true TE is identical to the assumed effect, and although there is a variation from study to study, the true TE is negative in only 2.5% of the trials. We call this scenario the optimistic scenario. Assuming a normal distribution of true TE, it can be expressed as

\[
\theta_i \sim N(\theta_A, (0.5\theta_A)^2)
\]

Next we consider a scenario where on average the true TE is half of the assumed effect. Keeping the assumption on normality and the degree of variation of the true TE, this means that the true effect is negative in 15.8% of all trials, and at least as large as the assumed effect in another 15.8%. This moderate scenario is given by

\[
\theta_i \sim N(0.5\theta_A, (0.5\theta_A)^2)
\]

Last, we consider a scenario where on average there is no effect, and in only 2.5% of all trials the true effect is at least as large as the assumed effect. This scenario named pessimistic scenario is given by

\[
\theta_i \sim N(0, (0.5\theta_A)^2)
\]

Table 2 summarizes the main properties of the three scenarios considered.

### 2.6. Scenarios for outcomes

It is an implicit assumption in our considerations that the values of \( E \) and \( P \) are mainly determined by the structure of our framework described so far, as long as \( \phi \) and \( \phi \) are based on asymptotically most powerful tests. They should not depend on the concrete choice of outcome scales, nuisance parameters of the outcome distribution, and inference procedures. However, to compute \( E \) and \( P \), a concrete choice has to be made.

In this paper, we present the results for the case of a binary outcome with the risk difference as the effect measure, i.e., \( \theta_{gi} = \pi^E_{gi} - \pi^S_{gi} \) with \( \pi^S_{gi} \) and \( \pi^E_{gi} \) denoting the probability of a success for

| Scenario | Distribution | Implication for distribution of true study effect |
|----------|--------------|--------------------------------------------------|
| Optimistic | \( \theta_i \sim N(\theta_A, (0.5\theta_A)^2) \) | Negative in only 2.5% of the trials |
| Moderate | \( \theta_i \sim N(0.5\theta_A, (0.5\theta_A)^2) \) | Negative in 15.8% of all trials, and at least as large as the assumed effect in another 15.8% |
| Pessimistic | \( \theta_i \sim N(0, (0.5\theta_A)^2) \) | At least as large as the assumed effect in only 2.5% of all trials |
patient \( i \) in subgroup \( g \) in study \( s \), treated by either the standard treatment or the experimental treatment, respectively. \( \pi_{sgi}^S \) is chosen as the constant value 0.4, implying that there is no association of the individual TEs with the prognosis under the standard treatment. The decision rule \( \phi \) at the overall level is based on the \( \chi^2 \) test, confidence intervals for the risk difference are based on Newcombes method No. 10 (Newcombe, 1998), and interaction tests are based on logistic regression. The power calculation of each single study is based on assumed success rates of 0.6 and 0.4 in the two treatment groups, i.e., \( \theta_A = 0.2 \). Consequently, a 90% power requires a sample size of \( N = 280 \) for each study. As we assume a normal distribution for \( \theta_{sgi} \), we cannot avoid that \( \pi_{sgi}^E \) is out of the interval \([0,1]\) for some patients, and hence we change \( \theta_{sgi} \) to \( \max(0, \min(1, \pi_{sgi}^S)) - \max(0, \min(1, \pi_{sgi}^E)) \) in computing the values of \( E \) and \( P \).

We focused on the risk difference in the main analysis, as it gives us a simple interpretation of \( E \). To check our implicit assumption, we considered six further scenarios in a sensitivity analysis, allowing an association between prognosis and individual TEs, a different overall prevalence, the use of the log odds ratio as effect measure and continuous outcomes with Cohen’s \( d \) as TE measures. Details are outlined in Appendix II and results are shown in Figure 5–7 in Appendix III of Supplementary Material.

3. Results

We start with illustrating the use of \( E \) and \( P \) being applied to ordinary clinical trials with no subgroup analyses. Fig. 2 shows the histograms of the individual true TEs in the three scenarios introduced in Section 2.5 as well as those of patients in studies resulting in a decision on superiority. In the optimistic scenario, the vast majority of trials results in such a decision, and consequently the value of \( E \) is close to the assumed TE of 0.2. The fraction of patients suffering from suggesting an inferior treatment is very small, i.e., within 3%. In the moderate scenario, less than 50% of the trials result in a positive decision, and consequently the value of \( E \) is less than half of the assumed effect, and \( P \) increases to 7%. In the pessimistic scenario, only very few trials result in a positive decision, and consequently the value of \( E \) is less than half of the assumed effect, and \( P \) increases to 7%. In the pessimistic scenario, only very few trials result in a positive decision. As in these few trials the true TE tends to be positive, we have still a small overall gain indicated by a value of 0.02 for \( E \). Also the control of \( P \) is still functioning to some degree, with limiting it to 14%.

Plots of \( P \) and \( E \) using different decision rules for the optimistic, moderate, and pessimistic scenarios with \( \tau = 0.5 \) and \( \tau = 1 \) and five values of \( R^2 \) are shown in Fig. 3. In these plots, we connected the points belonging to the two families of superiority testing \( \phi^{S,a} \) and inferiority testing.
\( R^2 = 0.0 \)  
\( R^2 = 0.125 \)  
\( R^2 = 0.25 \)  
\( R^2 = 0.375 \)  
\( R^2 = 0.5 \)

Figure 3. Plot of \( E \) vs. \( P \) using different decision rules for three scenarios with \( \tau = 0.5/1 \) and five different values of \( R^2 \). The results for the decision rules \( \phi^{\alpha} \) and \( \phi^N \) are connected by a line starting with \( \phi^{\alpha} \), i.e., no subgroup analysis, in the upper right corner and ending with \( \phi^N \) (marked by a circle). The two points marked in between correspond to \( \phi^F \) (marked by a diamond) and \( \phi^N \). The cross and plus correspond to \( \phi^{0.05} \) and \( \phi^{0.05} \) respectively. The filled and hollow triangles correspond to \( \phi^{0.75} \) and \( \phi^{0.75} \). Note that y- and x-scales vary from plot to plot.

\( \phi^{\alpha} \), such that we have a route from the case of no subgroup analysis over the case of comparing the estimate with 0, to the case of superiority testing at levels of 20% and 5%. Although the magnitude of \( P \) and \( E \) differs across the three scenarios and depends on \( \tau \), the general patterns observed are very similar. There are several different important aspects we can read out of Fig. 3. We first start with looking at rather liberal approaches, i.e., the family \( \phi^{\alpha} \) with \( \alpha \in (0.0, 1.0) \), starting with \( \alpha = 0.0 \), i.e., the case of no subgroup analysis \( (\phi^N) \), and moving to \( \alpha = 1.0 \), corresponding to the case of only comparing the estimate with 0(\( \phi^F \)). In Fig. 3, this means to start from the point in the upper right corner, and to follow the lines until the diamonds. For both values of \( \tau \) and all choices of \( R^2 \) in all scenarios, we can observe a reduction in \( P \) and a slight increase or decrease in \( E \). The reduction in \( P \) is increasing when \( R^2 \) and \( \tau \) increase, reflecting that the overall fraction of patients with a negative TE increases. The decrease in \( E \) is largest for \( R^2 = 0 \), as we then perform subgroup analyses without any need. An increase in \( E \) can be observed if \( \tau = 1 \) and for large values of \( R^2 \), i.e., we remove more patients with negative TE than with a positive TE using subgroup analysis. The %E and %P columns in the middle part of Table 3 quantifies the changes observed for \( E \) and \( P \) in Fig. 3, comparing \( \phi^E \) with \( \phi^N \). The decrease in \( P \) is less than 0.5% if \( R^2 = 0 \), between 3% and 11% if \( R^2 = 0.25 \), and between 11% and 27% if \( R^2 = 0.5 \). The decrease in \( E \) is always less than 0.75%, and in the case of substantial heterogeneity of TE, i.e., \( \tau, R^2 \) are large, the increase maybe up to 9%.

Then we continue with the more stringent approaches, i.e., the family \( \phi^{\alpha, \tau} \) with \( \alpha \in (0.05, 1.0) \), starting with \( \alpha = 1.0 \), i.e., the case of comparing the estimate with 0, and moving to \( \alpha = 0.05 \), i.e., significance testing at 5% level in each subgroup. This means in Fig. 3, we follow the lines from the diamonds to the circles at the bottom. The points we came across in between refer to superiority testing at 20%. We can always observe that \( P \) and \( E \) are both decreasing, i.e., we reduce the gain, as well as the
fraction of patients being recommended an inferior treatment. However, the reductions are now much more pronounced comparing to moving from $\phi^N$ to $\phi^E$ only. The %$E$ and %$P$ columns in the right part of Table 3 quantifies these reductions. The reduction in $P$ is in the range of 9–29% if $R^2 = 0$, and between 64% and 73% if $R^2 = 0.5$. The reduction in $E$ is in the range of 16% - 40% if $R^2 = 0$, and between 6% and 26% if $R^2 = 0.5$. So whether superiority testing at 5% is a good idea or not, depends on the degree of heterogeneity and how we balance the undesirable reduction in $E$ against the desirable reduction in $P$.

The next aspect depicted in Fig. 3 is the value of using interaction tests as gatekeepers, i.e., specific members of the family $\phi^{1,5,\alpha}$. Two instances of this strategy are considered: performing subgroup tests at $\alpha = 5\%$ if the interaction test was significant at either $\delta = 5\%$ or $\delta = 15\%$. The results are marked by a cross or plus, respectively. We can immediately see that the cross and plus are never above the line in the corresponding color, indicating that we can always find such a significance level that superiority testing yields a larger value of $E$ and a smaller value of $P$ simultaneously. This suggests that this strategy does not offer a general advantage. Similarly, Fig. 3 depicts the value of requiring to preserve a pre-specified fraction of the overall estimate, i.e., specific members of family $\phi^{E,\gamma}$, represented by the filled and hollow triangles, referring to a fraction of $\gamma = 0.5$ and $\gamma = 0.75$, respectively. Again, these symbols are never above the corresponding lines, hence, we cannot conclude a general advantage comparing to superiority testing. However, they tend to be closer to the curves than the symbols referring to the interaction gatekeeper strategy.

So far, we only considered the case of 2 subgroups. The upper part of Fig. 4 shows the corresponding plots for the moderate scenario with $\tau = 0.5$, but comparing the case of 2, 3, or 4 subgroups. We observe patterns similar to Fig. 3. The reduction in $P$ does not seem to depend on the number of groups, but the degree of reducing $E$ increases with the number of subgroups increasing. For example, for $R^2 = 0.25$ and superiority testing at 5% level, we observe a reduction by 24.8% in the case of 2 groups, by 39.1% in the case of 3 groups, and 49.6% in the case of 4 groups. Note that the rule of comparing the estimate with 0 is still associated with a moderate loss in $E$. With an increasing number of subgroups, the interaction gatekeeper approach tends to become much worse than superiority testing, and the discrepancy to superiority testing also increases for the fraction of estimate approach.

Finally, we consider the effect of increasing the sample size $N$ in the lower part of Fig. 4, simulating the situation that we can assess a benefit based on two or four studies, which are homogeneous enough to allow to be pooled together. Again, we observe the same patterns. Not surprisingly, the increase of power leads to an increase in $E$, and applying superiority testing at the

| Scenario  | $R^2$ | $\phi^N$ | $\phi^E$ | $\phi^{0.05}$ |
|-----------|-------|----------|----------|-------------|
| $\tau = 0.5$ | 0 | 0.1790 0.0325 | 0.1788 0.15 | 0.0324 0.34 | 0.1495 16.5 | 0.0232 28.59 |
| Optimistic | 0.25 | 0.1788 0.0326 | 0.1782 0.33 | 0.0310 5.04 | 0.1507 15.75 | 0.0149 54.35 |
| Moderate | 0.5 | 0.1792 0.0329 | 0.1785 0.42 | 0.0275 16.40 | 0.1518 15.28 | 0.0088 73.23 |
| $\tau = 0.5$ | 0 | 0.0784 0.0740 | 0.0781 0.29 | 0.0737 0.49 | 0.0562 28.22 | 0.0562 24.04 |
| Optimistic | 0.25 | 0.0788 0.0737 | 0.0783 0.65 | 0.0697 5.39 | 0.0593 24.75 | 0.0371 49.68 |
| Moderate | 0.5 | 0.0785 0.0726 | 0.0781 0.55 | 0.0619 14.76 | 0.0612 22.07 | 0.0225 69.08 |
| $\tau = 0.5$ | 0 | 0.0207 0.1386 | 0.0206 0.50 | 0.1386 0.001 | 0.0125 39.51 | 0.1129 18.52 |
| Pessimistic | 0.25 | 0.0207 0.1401 | 0.0205 0.74 | 0.1348 3.81 | 0.0143 30.95 | 0.0735 47.51 |
| Moderate | 0.5 | 0.0206 0.1377 | 0.0206 0.14 | 0.1218 11.53 | 0.0152 26.41 | 0.0482 64.98 |
| $\tau = 1$ | 0 | 0.1761 0.1389 | 0.1760 0.09 | 0.1387 0.12 | 0.1466 16.75 | 0.1211 12.77 |
| Optimistic | 0.25 | 0.1765 0.1390 | 0.1763 0.14 | 0.1250 10.05 | 0.1522 13.78 | 0.0791 43.07 |
| Moderate | 0.5 | 0.1760 0.1383 | 0.1779 $-1.07$ | 0.1019 26.36 | 0.1580 10.23 | 0.0465 66.40 |
| $\tau = 1$ | 0 | 0.0775 0.2037 | 0.0773 0.26 | 0.2034 0.15 | 0.0552 28.79 | 0.1806 11.34 |
| Moderate | 0.25 | 0.0780 0.2035 | 0.0783 $-0.37$ | 0.1817 10.72 | 0.0641 17.85 | 0.1157 43.13 |
| $\tau = 1$ | 0 | 0.0777 0.2044 | 0.0806 $-3.68$ | 0.1501 26.58 | 0.0697 10.37 | 0.0692 66.16 |
| Pessimistic | 0.25 | 0.0205 0.2664 | 0.0204 0.29 | 0.2661 0.13 | 0.0124 39.75 | 0.2422 9.08 |
| Moderate | 0.5 | 0.0207 0.2681 | 0.0212 $-2.55$ | 0.2388 10.94 | 0.0168 18.87 | 0.1531 42.90 |
| $\tau = 0.5$ | 0 | 0.0204 0.2680 | 0.0223 $-9.11$ | 0.1999 25.41 | 0.0192 6.01 | 0.0950 64.55 |
A level of 5% in two subgroups implies a level of E similar to no subgroup analysis in a study of half the size. In contrast, the impact on P is less pronounced, but we can observe a slight increase with larger sample sizes. The discrepancy between superiority testing and the interaction gatekeeper approach seems to decrease slightly when increasing the sample size, whereas the discrepancy to fraction of estimate tends to increase slightly. Selected values corresponding to Fig. 4 can be found in Table 4 in Appendix III of Supplementary Material.

4. Discussion

In this paper, we propose a framework that allows us to compare different subgroup analysis strategies applied to clinical trials which have demonstrated an overall effect.

As expected, using subgroup analysis can help reducing the number of patients who are suffering from an incorrect recommendation. The stricter the criterion we employ for performing subgroup analysis, the smaller the fraction of patients with incorrect recommendation among all with such a recommendation. However, we often have to pay the price for reducing the overall gain E while reducing P. If there is no subgroup variation, we have the maximal reduction in E and the minimal reduction in P. Even in the case of substantial subgroup variation, e.g., if the subgroups explain at least 50% of overall variation in TEs from patient to patient, strict decision rules like superiority testing at the 5% level can imply a substantial reduction of E, in particular when more than 2 subgroups are considered.

The simple decision rule of requiring a positive TE estimate in a subgroup implies only a small loss in E in the worst case, and if both large subgroup variation and large individual variation exist,

Figure 4. Plot of E vs. P for different number of subgroups K (3 plots in the upper part) and using different overall sample sizes N (3 plots in the lower part) for the moderate scenario and \( \tau = 0.5 \).
For the explanation of symbols see Fig. 3. Note that y- and x-scales vary from plot to plot.
we could even obtain a gain in \( E \). Simultaneously we always reduce \( P \) to a non-negligible degree in case of certain degree of subgroup variation. So here we have or are close to a “win–win” situation. Consequently, the use of superiority testing in subgroup analysis destroys this “win–win” situation. Any attempt to use some type of superiority testing in subgroup analysis destroys this “win–win” situation.

Interaction tests are often performed as a gatekeeper of subgroup analysis. However, we found in our framework that decision rules using the interaction gatekeeper approach generally do not offer any advantage. Similarly, requiring a certain fraction of the overall effect — a decision rule adopted by the Japanese guideline for global clinical trials — does not offer advantages either.

Although our framework is based on considering single trials, it also allows to throw some light on the situation of performing benefit assessments based on a meta-analysis. On a qualitative level, subgroup analysis seems to behave very similar in meta-analysis with respect to the relation between different decision rules regarding the impact on \( E \) and \( P \). On a quantitative level, we can reach higher levels of \( E \). Of course we have to be aware of the fact that the gain in \( E \) is probably due to larger sample sizes allowing to demonstrate smaller TEs, for which the clinical relevance might be questionable.

In interpreting results, some limitations of our framework have to be taken into account. First, we assume that \( \sigma_i^2 \) i.e., the inter-individual variation of the TEs, is independent of the true TE \( \theta_i \) of the trial. It seems to be more realistic to assume that \( \sigma_i^2 \) decreases if \( \theta_i \) approaches 0, as if a treatment is not effective on average, it is unlikely that there are large effects at the individual level. However, assuming proportionality between \( \sigma_i^2 \) and \( \theta_i \) is also too strong. Our choice of assuming no association between \( \sigma_i^2 \) and \( \theta_i \) can be regarded as a conservative choice, implying an overestimation of \( P \), as we assume more negative individual TEs than to be expected in reality.

Second, our framework assumes that decisions on making a treatment available for all patients are based on a single trial. This is usually not the case. For example, the FDA and EMA guidelines assume the existence of two pivotal studies (European Medicine Agencies, 2000; Food and Drug Administration, 1998). However, the example of the CAPRIE trial illustrates that sometimes decisions are based on one single, large trial. Furthermore, our considerations for the case of pooling two or four studies suggest that many considerations and conclusions will be similar if more than one study is available.

Third, our framework assumes that a proof of superiority implies a decision to make the treatment available for all patients, whereas a failure to do so implies no access to the treatment in future. The latter may be the case if we consider decisions on drug approval by regulatory agencies (Maggioni et al., 2007), but then the first assumption is not fulfilled, as a proven gain in efficacy has to be balanced with the safety profile of the new treatment. If we consider decisions on reimbursement instead of drug approval (Paget et al., 2011), then also the failure to demonstrate superiority may not imply that the drug is not available for all patients. For example, it may be available for those who are willing to pay by themselves. In Germany, a negative decision on an additional benefit in an health technology assessment assessment performed after drug approval implies that the drug may still be sold, but just at the price of the comparator, and it is up to the pharmaceutical companies to decide whether they accept this or remove the drug completely from the market.

Fourth, in our framework we neglect that treatments are developed for patient populations of different sizes. Our results can change if there is a correlation between true TEs and the size of the target population. Fifth, for simplicity, we made the strong assumption that all subgroups in a trial are of equal size, which may not be true in most clinical trials. Similarly, we considered only one partition of the study population into subgroups. In most trials, several factors like age, gender, or baseline characteristics may define different partitions, adding the additional complexity of patients with contradicting treatment recommendations. Sixth, we assume in our framework that TEs observable in clinical trials will also be
observed when the treatment later is applied as part of the standard care. Finally, we did not include all approaches of subgroup analyses in our consideration. However, our framework will allow to investigate other alternative approaches, e.g., Bayesian methods (Jones et al., 2011).

5. Conclusion

Subgroup analysis is a great temptation to improve the benefit assessment of new drugs by allowing to consider smaller relevant patient groups. However, there is a risk of overlooking superior new treatments due to insufficient power, which may decrease the overall efficiency of the clinical trial culture to improve the average outcome. The simple rule of comparing effect estimates instead of confidence intervals with 0 in subgroup analyses in trials with a significant overall effect may supply a good compromise. It allows us to reduce the fraction of patients suffering from being recommended a worse treatment without diminishing the overall gain achievable by the current clinical trial culture. More strict decision rules may be justified, if there is a priori a high likelihood for substantial subgroup variation, or if avoidance of incorrect treatment recommendations change at the individual level are given more weight than improvement of the average outcome.

Appendix I

Structure and technical details

We generate the observations from the centered normal distribution by generating observations from a normal distribution and then centering them to mean 0. As this decreases the variance, we use in the first step an appropriately increased variance. The centered normal distributions are employed at both the subgroup and individual patient level:

- $\theta_{sg} = \theta_s + \varepsilon_{sg}$ with $\varepsilon_{sg} = \varepsilon_s - \bar{\varepsilon}_s$ and $\varepsilon_{sg} \sim N\left(0, \sigma^2_{sg}\right)$
- $\theta_{sgi} = \theta_{sg} + \varepsilon_{sgi}$ with $\varepsilon_{sgi} = \varepsilon_{sgi} - \bar{\varepsilon}_{sgi}$ and $\varepsilon_{sgi} \sim N\left(0, \sigma^2_{sgi}\right)$

Appendix II

Further simulation scenarios

(1) The first additional scenario is identical to the one described in the paper, except that we chose $\pi_{sgi}^S = 0.1$ and assumed $\theta_A = 0.1$, which requires a sample size of $N = 572$. Results are presented in Figure 5 in Appendix III of supplementary materials.

(2) The second scenario differs from the one considered in the paper, only by $\pi_{sgi}^S = 0.5 - 0.5 \times \theta_{sgi}$. So here we allow an association of individual treatment effects with prognosis under standard treatment.

(3) Instead of risk difference, we consider the $\theta = \log(OR)$ as effect measure. Power calculation and test of overall effects are not changed. Two scenarios were considered with respect to the association between prognosis and individual treatment effect:

(a) $\pi_{sgi}^E = 0.4$ and $\pi_{sgi}^E = \logit^{-1}(\logit(0.4) + \theta_{sgi})$
(b) $\pi_{sgi}^S = \logit^{-1}(\logit(0.5) - 0.5 \times \theta_{sgi})$ and $\pi_{sgi}^S = \logit^{-1}(\logit(0.5) + 0.5 \times \theta_{sgi})$

Results of scenario (a) are shown in Figure 6 in Appendix III of supplementary materials. Note that in this scenario, the subgroup specific effect $\theta_{sgi}$ and the study effect $\theta_s$ do not refer to the group or study specific log(OR) values, but to the average of the individual log(OR) values.
Finally we consider continuous outcomes and Cohen’s $d$ as effect measure, i.e., $	heta = (\mu_E - \mu_S)/\sigma^2$. The t-test was used as overall test. CIs were t-test based, and a regression model was used to assess interactions. The assumed effect was 0.2, we could assume $\sigma^2 = 1$ implying a sample size of 1052. The data was generated as:

$$Y_{sgi} = \mu_{sgi} + \varepsilon_{sgi} \text{ with } \varepsilon_{sgi} \sim N\left(0, 1 - \sigma^2_{sgi}\right)$$

(11)

where

$$\mu_{sgi} = \begin{cases} \mu^S_{sgi} & \text{if } T = 0 \\ \mu^E_{sgi} & \text{if } T = 1 \end{cases}$$

and

$$\sigma^2_{sgi} = \begin{cases} \sigma^2_S & \text{if } T = 0 \\ \sigma^2_E & \text{if } T = 1 \end{cases}$$

Two scenarios were assumed in generating data with respect to the association between prognosis and individual treatment effects

(a) $\mu^S_{sgi} = 0$ and $\mu^E_{sgi} = \theta_{sgi}$, implying $\sigma^2_S = 0$ and $\sigma^2_E = \sigma^2_I$

(b) $\mu^S_{sgi} = -0.5\theta_{sgi}$ and $\mu^E_{sgi} = 0.5\theta_{sgi}$, implying $\sigma^2_S = \frac{1}{4}\sigma^2_I$ and $\sigma^2_E = \frac{1}{4}\sigma^2_I$

Results of scenario(a) are shown in Figure 7 in Appendix III of supplementary materials.

Software

Software in the form of Stata code, together with sample simulation data sets is available on request from the corresponding author (E-mail: sun@imbi.uni-freiburg.de).

Acknowledgments

The authors are grateful to Frank Bretz, Marc Buyse, Martin Schumacher, and two referees for providing valuable and helpful comments. We also thank Kim Harris for English editing.

Funding

This work was supported by funding from the European Community’s Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES (“Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews”; www.mediases-itn.eu) with the Grant Agreement Number 290025.

References

Bender, R., Koch, A., Skipka, G., Kaiser, T., Lange, S. (2010). No inconsistent trial assessments by NICE and IQWiG: different assessment goals lead to different assessment results regarding subgroup analyses. *Journal of Clinical Epidemiology* 63:1305–1307.

Brookes, S., Whitely, E., Egger, M., Smith, G., Mulheran, P., Peters, T. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology* 57:229–236.

CAPRIE Steering Committee. (1996). A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). *The Lancet* 348:1329–1339.

Dent, L., Raftery, J. (2011). Treatment success in pragmatic randomised controlled trials: a review of trials funded by the UK Health Technology Assessment programme. *Trials* 12:Article 109.

Djulbegovic, B., Kumar, A., Miladinovic, B., Reljic, T., Galeb, S., Mhaskar, A., Mhaskar, R., Hozo, I., Tu, D., Stanton, H., et al. (2013). Treatment success in cancer: industry compared to publicly sponsored randomized controlled trials. *Plos One* 8:Article e58711.

Djulbegovic, B., Kumar, A., Soares, H., Hozo, I., Bepler, G., Clarke, M., Bennett, C. (2008). Treatment success in cancer. *Archives of Internal Medicine* 168:632–642.

European Medicines Agency. (2000). Points to consider on validity and interpretation of meta-analyses, and one pivotal study. Released in October 2000: [http://www.emAdda.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003659.pdf](http://www.emAdda.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003659.pdf).
European Medicines Agency. (2010). Concept paper on the need for a guideline on the use of subgroup analyses in randomised controlled trials. Released on 22-04-2010: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/05/WC500090116.pdf.

European Medicines Agency. (2014). Draft guideline on the investigation of subgroups in confirmatory clinical trials. Released in January 2014: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf.

FDA Advisory Committee. (1997). 82nd meeting of the cardiovascular and renal drugs committee. Released on 24-10-1997: http://www.fda.gov/ohrms/dockets/ac/97/transcpt/3338t2.pdf.

Food and Drug Administration. (1998). Guidance for industry: Providing clinical evidence of effectiveness for human drug and biological products. Released in May 1998: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm078749.pdf.

Hasford, J., Bramlage, P., Koch, G., Lehmacher, W., Einhaupl, K., Rothwell, P. (2010). Inconsistent trial assessments by the National Institute for Health and Clinical Excellence and IQWiG: Standards for the performance and interpretation of subgroup analyses are needed. Journal of Clinical Epidemiology 63:1298–1304.

ICH Steering Committee. (1999). Statistical principles for clinical trials (ICH E9). Statistics in Medicine 18:1905–1942.

Ikeda, K., Bretz, F. (2010). Sample size and proportion of Japanese patients in multiregional trials. Pharmaceutical Statistics 9:207–216.

IQWiG. (2006). Final report plan for the evaluation: “Clopidogrel versus acetyl-salicylic acid in the secondary prevention of vascular events”. Released in January 2006: http://www.iqwig.de/download/A04-1A_Abschlussbericht_Clopidogrel_versus_ASS_in_der_Sekundaerprophylaxe.pdf.

Jones, H., Ohlssen, D., Neuenschwander, B., Racine, A., Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. Clinical Trials 8:129-143.

JPFSB (2007). Japanese Pharmaceutical and Food Safety Bureau: Basic principles on global clinical trials. Released on 28-09-2007: http://www.pmda.go.jp/operations/notice/2007/file/0928010-e.pdf.

Kumar, A., Soares, H., Wells, R., Clarke, M., Hozo, L., Bleyer, A., Reaman, G., Chalmers, I., Djulbegovic, B. (2005). Are experimental treatments for cancer in children superior to established treatments? Observational study of randomised controlled trials by the Children’s Oncology Group. British Medical Journal 331:1295–1298B.

Maggioni, A., Darne, B., Atard, D., Abadie, E., Pitt, B., Zannad, F. (2007). FDA and CPMP rulings on subgroup analyses. Issues in Drug Development 107:97–102.

Newcombe, R. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. Statistics in Medicine 17:873–890.

NICE. (2005). Clopidogrel and modified-release dipyridamole in the prevention of occlusive vascular events. Released in May 2005: www.nice.org.uk/nicemedia/pdf/TA090guidance.pdf.

Paget, M., Chuang-Stein, C., Fletcher, C., Reidd, C. (2011). Subgroup analyses of clinical effectiveness to support health technology assessments. Pharmaceutical Statistics 10:532–538.

Pocock, S., Assmann, S., Enos, L., Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine 21:2917–2930.

Pocock, S., Calvo, G., Marrugat, J., Prasad, K., Tavazzi, L., Wallentin, L., Zannad, F., Garcia, A. (2013). International differences in treatment effect: do they really exist and why? European Heart Journal 34:1846–1852.

Yusuf, S., Wittes, J., Probstfield, J., Tyroler, H. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. Journal of The American Medical Association 266:93–98.