Chapter 15
Subgroup Analysis: A View from Industry

Oliver N. Keene and Daniel J. Bratton

Abstract Subgroup analysis in clinical trials for regulatory and reimbursement purposes can be confirmatory or exploratory in nature. Confirmatory subgroup analysis requires pre-specification of the proposed analysis and appropriate control of the type I error rate. Exploratory subgroup analysis is a feature of Phase III clinical trials. Examination of the results by sex, age and race is required by FDA and submissions for regulatory approval typically involve numerous further analyses by baseline characteristics such as disease severity. For efficacy these exploratory analyses are often directed at providing reassurance that the overall estimated treatment effect translates into benefit for each of the subgroups and for safety to investigate the existence of signals in more vulnerable subgroups. For reimbursement purposes, extensive analysis is required to try to identify those groups experiencing most benefit and for whom the medicine is therefore most cost-effective.

Exploratory subgroup analyses present a major challenge in interpretation due to the large number of subgroups examined. Effect sizes can vary largely from the overall treatment effect estimate and even be in opposite directions due to chance alone. The commonly used statistical methods to assess consistency of effect all have limitations. There is an important role for statistical modelling and an increasing interest from industry in Bayesian shrinkage techniques which balance emphasis on a specific observed differential subgroup effect with the overall treatment effect.

When planning and designing confirmatory trials of new medicines, discussion and agreement with regulatory and reimbursement authorities on the population is exceptionally valuable. Pre-identification of a small number of important biologically plausible subgroups which require exploration is helpful for interpretation.

O. N. Keene (✉) · D. J. Bratton
GlaxoSmithKline Research and Development, Middlesex, UK
e-mail: oliver.n.keene@gsk.com

N. Ting et al. (eds.), *Design and Analysis of Subgroups with Biopharmaceutical Applications*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-030-40105-4_15

15.1 Introduction

The classic design of a company-sponsored late-stage trial is directed at providing a single overall estimate of the effect of a medicine on the primary endpoint. Many important stakeholders find a single summary of response on an endpoint to be incomplete. Patients want to know if this average effect will apply to them with their own individual set of baseline characteristics which will vary among patients studied in the clinical trial. Physicians are concerned with identifying those patients for whom the medicine will be more effective or less effective. Payers only want to pay for a medicine for patient groups where the medicine is cost-effective.

The need for subgroup analyses is therefore unavoidable for late stage clinical trials performed by the pharmaceutical industry. They are regularly requested by practising physicians seeking to understand the results of the trial in the context of the diversity of the patients who consult them.

In a regulatory setting, the FDA require summaries of efficacy and safety by demographic subgroups (FDA 2015) and for a multi-regional trial an evaluation of consistency of treatment effects across regions is required by ICH E17 (ICH 2017). In a reimbursement setting, the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany requires analysis by sex, age, country and disease severity for all patient relevant endpoints, including safety endpoints as well as efficacy endpoints (IQWiG 2017). These requirements are independent of an a priori expectation that a particular subgroup will experience a different treatment effect. As well as these mandated subgroups, further subgroup analyses are also frequently requested by regulatory and reimbursement agencies to assess consistency of treatment effects.

In the next sections we start by defining what is meant by a subgroup effect. We then review the key issue of multiplicity. Later sections describe analysis methods that go beyond the simple approaches of separate analysis of subgroups according to a specific characteristic and interaction tests.

15.2 Defining a Subgroup Effect

It is important to define what is meant by a subgroup effect as this terminology can have different interpretations. Subgroups can be dichotomous (e.g. male/female), categorical (e.g. region), ordered categorical (e.g. disease score at baseline) or based on a continuous measure (e.g. age). For subgroups defined by a continuous measure, patients are often categorised based on values lying within specific cut-points. A more powerful method of evaluation is often to retain the continuous scale and use a modelling approach (this is discussed later in the chapter).

Subgroup effects considered in this chapter are defined by baseline characteristics measured prior to treatment. Analysis based on differentiating patients according to a post-randomisation measurement can be misleading, because a particular treatment effect may influence classification to the subgroup (Yusuf et al. 1991).

Table 15.1 Illustrative example of importance of scale of measurement for subgroup effects

Baseline event rate	Placebo event rate	Active event rate	Absolute reduction	Percentage reduction (%)
0	0.8	0.6	0.2	25
1	1.2	0.9	0.3	25
2 or more	2.0	1.5	0.5	25

There are two keys aspects to describing a subgroup effect for a typical phase III superiority study. Firstly, the baseline characteristic may affect the outcome regardless of treatment and therefore be a prognostic variable. In modelling terms, this would be a main effect. For example, severe patients may have poorer outcomes than milder patients. Secondly, a baseline characteristic may influence the effect of active treatment compared to placebo and therefore be a predictive variable. In modelling terms this would be an interaction of the treatment effect with the variable. The same covariate can be both prognostic and predictive; it is examination of potential predictive variables that is the focus of this chapter.

Importantly, whether a differential treatment effect exists may depend on the scale of measurement used (Keene 1995). For instance, consider the example below shown in Table 15.1. The outcome variable is the number of events during treatment, and this has been split according to number of events in the previous year. For those with a baseline event rate of 0 per year, the event rate after randomisation is 0.8 on placebo compared to 0.6 on treatment, a reduction of 25%. The same percentage reduction of 25% applies to those with two or more events in the previous year. However, some may consider absolute reductions as more clinically relevant; these are very different, a reduction of 0.5 events/year for this group compared to 0.2 events/year for the group with 0 events at baseline. Model based analysis of event rates such as the negative binomial model (Keene et al. 2007) express treatment effects in terms of relative reductions and therefore there would be no statistical interaction. However, for a payer, there may be more willingness to fund a medicine that reduces event rates by 0.5 events a year than one that reduces events by 0.2 events.

15.3 Multiplicity in Subgroup Analysis

The major difficulty when interpreting subgroup analysis is that subgroup differences in treatment effect can arise by chance and it is exceptionally hard to identify what is a true difference. While there is a general acknowledgement that results from small subgroups are unreliable, unfortunately results from analyses of larger subgroups of patients are often interpreted as the true results for that group of patients, ignoring the fact that it is likely that some groups will show bigger or smaller differences simply by chance. While multiplicity issues can also arise in

clinical trials from other sources such as multiple endpoints, the issue is particularly difficult for subgroups.

In the classic illustration of the problem, the ISIS-2 authors (ISIS 1988) examined the outcome by astrological birth sign. While the overall results showed an impressively positive effect for aspirin on mortality, for patients born under Gemini or Libra there was a small observed increase in mortality.

For trials performed by the pharmaceutical industry, prior specification of a subgroup, combined with an appropriate strategy to strongly control the type I error rate is required if a claim of efficacy in a subgroup is to be approved (EMA 2002). If the effect of a treatment is expected to be stronger in a subgroup compared to the complimentary subgroup, then studying this subgroup alone is an option. However, where the subgroup is defined by a biomarker, there is a desire from regulatory authorities to understand the effect in both the biomarker positive and biomarker negative patients. The FDA guideline on enrichment designs (FDA 2019) suggests that the type I error rate for the study be shared between a test conducted using only the enriched subpopulation and a test conducted using the entire population. In this case, it will be beneficial to increase sample size in the group where greater efficacy is expected. Simple strategies for this sharing include Bonferroni adjustment of p-values or hierarchical testing but increases in power can be obtained using strategies that take advantage of the correlation between the test statistics for analysis of a subgroup and analysis of the whole population (Song and Chi 2007).

Historically, there have been concerns about inferring efficacy in a post-hoc subgroup in trials where the overall effect was not positive. However, the current emphasis in regulatory and reimbursement submissions is on showing that specific subgroups derive benefit from the medicine in the presence of a positive effect overall. Regulators seek assurance that effects are consistent across subgroups and payers seek to restrict access to medicines to those groups where the benefit is strongest.

Li et al. (2007) investigated the probability of observing negative subgroup results when the treatment effect is positive and homogeneous across subgroups. Negative here is defined as an effect size in the opposite direction to the overall result. They show that if a trial with 90% power to detect an overall effect and total sample size of 338 is divided into five equally sized subgroups, the probability of observing at least one negative subgroup result is 32%. Each subgroup in this case has more than 65 patients.

The number of different subgroups that are typically examined in a confirmatory clinical trial of a new medicine is extensive and this can create challenges in interpretation. For an integrated summary of effectiveness, the FDA guideline (FDA 2015) includes the following list of subpopulations to be considered: age, sex, race, disease severity, prior treatment, concomitant illness, concomitant drugs, alcohol, tobacco, body weight and renal or hepatic functional impairment. While some of these subpopulations may not be applicable to a specific medicine, most will be

and there will be a perceived need to split the data into subgroups according to multiple criteria. The greater the number of subgroup analyses performed and the smaller the resulting subgroups are, the higher chance that there will be subgroups with seemingly no benefit or potential harm from treatment. This issue of selection bias is recognised in the European regulatory guideline on subgroup analysis (EMA 2019) which states: "Not only might the play of chance impact the estimated effect, but it is tempting to focus on subgroups with extreme effects".

For submission for reimbursement in Germany, the Institute for Quality and Efficiency in Health Care (IQWiG) requires analysis by sex, age, country and disease severity for all patient relevant endpoints, including safety endpoints as well as efficacy endpoints (IQWiG 2017). These analyses are usually performed in the population for whom reimbursement is sought, often already a subgroup of the trial population. This requirement can typically lead to an excessively large number of subgroup analyses (e.g. 5 characteristics $\times$ 20 endpoints $=$ 100 subgroup analyses) and can involve very small sample size in some analyses. The credibility of the analysis produced for IQWIG submissions has therefore been questioned (Ruof et al. 2014) and the value of this exhaustive exercise in the determination of the cost-effectiveness of the medicine is unclear.

When assessing whether observed differences across levels of a subgroup represent a true difference, it is possible to use checklists such as that provided by Sun et al. (2010). In practice, discussion often focuses on biological rationale (Hemmings 2014; Pocock et al. 2002). Unfortunately, biological plausibility is a somewhat elusive concept as most subgroup analyses have a degree of plausibility and therefore it is helpful to plan subgroup analysis in advance of unblinding of the trial. One possibility is to divide proposals for subgroup analysis into whether (a) a differential effect is anticipated, (b) a differential effect is biologically plausible but not anticipated and (c) observed differential effects are hypothesis generating (Dane et al. 2019). The weight given to the observed findings could then depend on which category the subgroup analysis was assigned to as well as the overall number of subgroup analyses performed.

Replication across endpoints and across two or more trials strengthens the support for a hypothesis of a different effect in a specific subgroup. In particular because of regression to the mean, treatment effects from exploratory subgroup analyses that show the biggest differential often cannot be reproduced.

15.4 Statistical Methods

The next sections describe commonly used statistical methods for investigating exploratory subgroup effects. This is not an exhaustive list and other methods are available. The focus is on methods that explore treatment by a single covariate, which is a problem in many practical cases in analysis of data from clinical trials.

15.4.1 Separate Analysis by Subgroup

Separate analysis by subgroup can be performed by either using an entirely separate analysis for the specific subgroup or via a model of the complete data with a treatment interaction term for the subgroup under investigation.

Graphical representation of subgroup analyses is a key component in facilitating the interpretation of subgroup analyses. Forest plots, displaying treatment effect estimates for each subgroup along with the associated confidence interval, is one of the most common displays used.

Interpretation of such forest plots however is not straightforward. For example, it is not possible to draw valid inferences about consistency of effect by comparing the individual subgroup p-values or by assessing whether the CIs in the forest plot cross the line of no difference. A significant difference in one subgroup but not the other is not necessarily evidence of a significant difference between the subgroups.

When performing subgroup analysis, it is common to classify a continuous variable such as age into categories and to analyse each subgroup separately. A key choice then is the number and location of the cut-points used to define the categories. Usually these might be dictated by clinical relevance. For example, it is often necessary to define body mass index (BMI) subgroups by <18.5, $18.5–<25$, $25–<30$ and ≥ 30 kg/m^2 as these are commonly used in clinical practice to identify underweight, normal, overweight and obese patients respectively. Where possible, it is helpful to state the cut-points prior to unblinding as different choices of cut-points can result in different estimates of treatment effect (Royston et al. 2006).

However, pre-specifying cut-points is not without issues. In some cases, there might be insufficient data in a particular pre-defined subgroup to allow estimation of a treatment effect. In such cases the subgroup could be combined with a neighbouring group but then the analysis can lose some of its value in estimating treatment differences in groups of interest or even miss a true interaction. A potential solution to this problem is to define subgroups by quantiles of the observed covariate distribution (e.g. quartiles) to help ensure that there will be sufficient data within each subgroup. Although such an approach might help to identify associations between the treatment effect and the covariate, the chosen subgroups may not have an easy clinical interpretation.

15.4.2 Interaction Tests

The classical approach to assessing consistency of effects across subgroups is to perform an interaction test. The focus of interest here is the contrast between the effects in the different subgroups, rather than examining a specific subgroup in isolation. For a factor with multiple levels such as region, a global test of any difference across all categories can be performed or a test for a specific category

vs. the rest of the population. The ICH E9 guideline indicates in section 5.7 that interaction testing is the first step in undertaking subgroup analyses (ICH 1999).

However, in practice, simple significance tests for interaction are on their own of limited value when investigating subgroup differences. Firstly, as interaction tests are tests of significance, they have an associated fixed type I error rate. If this is fixed at 5%, then even if there are no true differences among subgroups, 5% of the tests will be expected to be significant suggesting a differential subgroup effect. Because of the low power of interaction tests, tests at the 10% or 20% level have been suggested (Hemmings 2014). In these cases, even more false positive results are to be expected.

Secondly, they have low power to detect heterogeneity. For example, in the simple case of a continuous endpoint with two equal sized groups, the variance of the interaction contrast is four times the variance of the overall treatment difference. This implies that only unlikely large interaction effects can be detected with any certainty.

Absence of statistically significant interactions does not imply consistency of the treatment effect in the studied population since absence of statistical significance cannot be taken to imply equality or consistency. To require only absence of statistical significance in an interaction test, or only directional consistency, would not be sufficiently sensitive filters to detect differences of potential interest.

The need to go beyond simple interaction tests is recognised in the CHMP guideline on subgroup analysis (EMA 2019) which states that "The sole reporting of an isolated p-value from a test for interaction is an inadequate basis for decision making". The guideline recommends including estimates of the size of the interaction contrasts with associated confidence intervals to show what differences a trial can reliably detect.

15.4.3 Stepwise Regression

When subgroups are assessed individually, the analysis does not account for potential imbalances in known effect-modifiers between groups. Multivariable analysis of an endpoint including various subgroups of interest and their interaction with treatment may be required to determine whether the effects observed within a subgroup are partially or wholly affected by other factors. In addition, a modelling approach allows for correlation between covariates instead of examining these in isolation.

Selection methods based on stepwise regression can be a useful exploratory approach for this to help determine the most influential factors on treatment effect. Various approaches are available for such an analysis, including forward, backward and stepwise selection (Royston and Sauerbrei 2009).

In backward selection, all subgroups of interest and their interactions with treatment are included in a model and the term with the largest p-value is removed if it is above a specified significance threshold (e.g. $p = 0.1$). The process continues

iteratively removing the term with the largest p-value above the significance threshold at successive steps until all remaining covariates are significant at this level. Alternatively, selection methods based on information criteria (e.g. Akaike's information criterion) or penalised likelihood could be used rather than p-values for the individual covariates. Main effects should only be removed if its interaction with treatment has also been removed at a previous step. Forward selection is essentially the opposite to this, with terms being added separately to a model and retaining that with the smallest p-value below some specified threshold for the next step of the process. This is repeated until no covariates are significant when added to the model. In this case interaction terms should only be added to the model if the corresponding main effect was added at a previous iteration. Stepwise selection is a combination of the two approaches, testing variables for inclusion or exclusion at each step and allowing previously included or excluded variables to be removed or reincluded respectively. Ideally all methods will result in the same final model, but this is not always the case.

Results from such models should be deemed to be exploratory in nature since the selection procedure will tend to lead to an over-estimation of the effects of the selected covariates and, as in the case of separate analyses of subgroups, type I error rate is not strictly controlled. However, they are a useful tool for hypothesis generation or building prediction models.

When control of type I error is required, then potential methods are reviewed in Dane et al. (2019), Ballarini et al. (2018) and Thomas and Bornkamp (2017). Dane at al describe resampling methods and Balletini et al use penalised regression with a Lasso-type penalty as a model selection and estimation technique. Thomas and Bornkamp include model averaging in addition to resampling and Lasso methods. However, absence of statistical significance does not imply that the effects are the same in each subgroup and in a response to the Dane et al. article, Hemmings and Koch (2019) argue that "power should be prioritised over type I error where the objective is to generate signals for further inspection".

15.4.4 Fractional Polynomial Modelling Approaches with Continuous Covariates

As stated above, it is common to analyse a continuous variable by classifying the variable into categories. Clear disadvantages of this approach are the loss of information (Altman and Royston 2006; Royston et al. 2006) and the assumption that patients close to a cut-point will have different responses when these are likely to be similar. While these subgroup analyses provide treatment effect estimates within a narrower range of the baseline covariate than in the overall study, they do not necessarily adequately estimate the effect of treatment for a particular value of that covariate which might be more useful to an individual patient.

A more informative approach is to create a statistical model of the outcome by treatment as a function of the covariate (Keene and Garrett 2014). For instance, the covariate of interest can be entered into the model as a continuous linear term along with its interaction with treatment. Such a model allows treatment differences to be estimated at particular values of the covariate of interest rather than in groups. A resulting plot of the estimated treatment difference versus the covariate can potentially show in more detail how the treatment effect varies over the range of the covariate than a forest plot of subgroup effects focusing on specific categories.

However, if the relationship between treatment efficacy and the covariate of interest is non-linear then a model where the prognostic and predictive effects of a covariate are represented by linear terms may fit the data poorly. For instance, if a treatment has lower efficacy in patients who are underweight (<18.5 kg/m^2) and overweight (≥ 25 kg/m^2) compared to patients in the 'normal' range, then the association between efficacy and BMI is non-linear. In such a case a linear model will miss such a result whereas the subgroup analysis is more likely to demonstrate this interaction. Other transformations of the covariate could be assessed (e.g. log transformation or adding a quadratic term) but while some transformations might fit the data better and more closely align with subgroup estimates, there may be more appropriate functions to use.

Fractional polynomial models (FPs) offer the flexibility to identify non-linear treatment-covariate interactions (Royston and Altman 1994). In the FP framework, various transformations of a covariate are assessed and the model which describes the data best is selected. Transformations of the covariate of interest X that are assessed are of the form X^p, where p is chosen from a set S of eight powers: $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Here $p = 0$ indicates a log transformation of X. Each transformed covariate is assessed individually and that which maximises the likelihood of the model is used to assess the treatment interaction (Royston and Sauerbrei 2004). The model can include other covariates, for instance those pre-specified in the primary analysis, or a multivariable fractional polynomial (MFP) algorithm can be applied prior to modelling the treatment interaction to determine the most influential prognostic factors for the outcome and their best fitting forms (Royston and Sauerbrei 2009).

An FP model containing a single transformation of the covariate X is referred to as an FP1 model. To increase the flexibility of the modelling procedure, two transformations of the covariate can be entered into the model using powers from the same set S, so the model contains terms X^p and X^q where $p, q \in S$. This is referred to as an FP2 model. If $p = q$ then this is referred to as a repeated-powers model and one of the terms is replaced by $X^p \log(X)$. Unlike an FP1 model, including two transformations of X allows non-monotonic functions to be fitted thus greatly increasing the flexibility of the modelling. Examples of FP1 and FP2 functions are shown in Fig. 15.1. More than two transformations of the covariate could be used, but such models do not greatly increase the flexibility of the modelling procedure over and above FP2 models, can greatly increase the time taken to find the best fitting model, and may lead to overfitting.

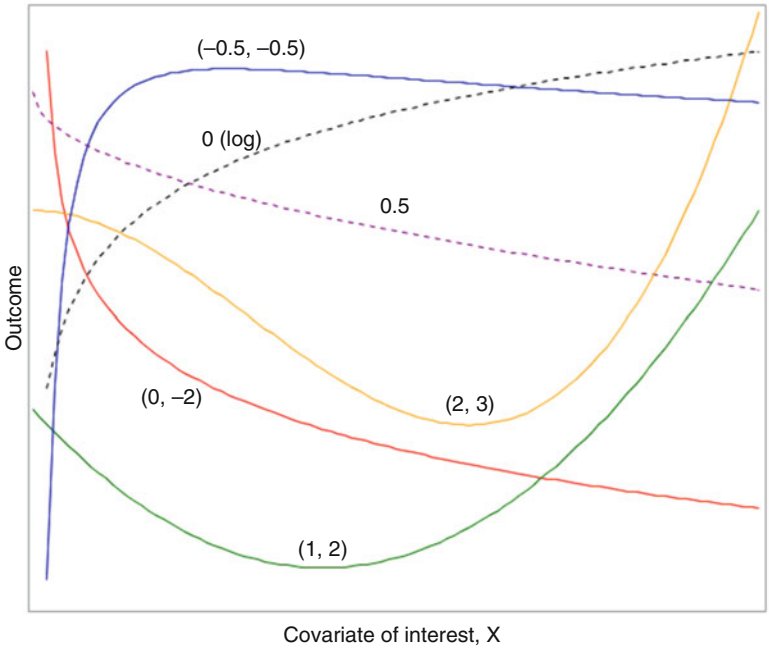

Fig. 15.1 Examples of FP1 (dashed lines) and FP2 (solid lines) functions

It is important to check the results of the FP modelling, particularly if it indicates a treatment-covariate interaction. Should there be an interaction, then this is also likely to be indicated by a subgroup analysis. Therefore, estimating treatment effects within a number of subgroups, for instance defined by quartiles or quintiles, can show whether there is agreement between the two approaches. Disagreement should be a signal of caution and investigated as it could be an artefact of the modelling—for instance due to influential outliers of the covariate which are less likely to affect a subgroup analysis.

Although FP modelling has several advantages over subgroup analysis, it is not without some potential pitfalls. FPs can behave strangely at the tails of the covariate, particularly close to 0 when negative powers are used. However, given that tails contain little data and that the CIs for the treatment effect line are likely to be wide, the plot of the treatment interaction can simply be truncated so that only the middle 90 or 95% of the distribution of the covariate are presented. There are also issues with scaling and ensuring that the covariate is strictly positive prior to modelling, but suitable solutions are available (Royston and Sauerbrei 2007).

An example of the value of a modelling approach is provided by the METREO and METREX trials of mepolizumab in patients with COPD (Pavord et al. 2017). These two randomised, placebo-controlled, double-blind, parallel group trials compared mepolizumab (100 mg in METREX, 100 or 300 mg in METREO) with placebo, given every 4 weeks for 52 weeks in patients with COPD who had a

history of moderate or severe exacerbations while taking inhaled triple maintenance therapy. The trials were funded by GlaxoSmithKline (ClinicalTrials.gov numbers: METREO: NCT02105961, METREX: NCT02105948). The primary variable was the rate of moderate/severe exacerbations and analysis was performed using a negative binomial generalised linear model with a log link function (Keene et al. 2007).

The key covariate of interest was the screening blood eosinophil count. A pre-specified meta-analysis of the two studies was performed to examine the result of the studies by subgroups defined by categories of screening blood eosinophil count and the results are shown in Fig. 15.2. The estimated exacerbation rate reduction in patients with a screening eosinophil count between 300 and <500 cells/μL is 18%, however, some patients are likely to fare better than others within this category and so it is not clear for example what the estimated treatment effect is for a patient with say an eosinophil count of 400 cells/μL. In addition, the subgroup analysis implies a cliff effect at the cut-offs whereby two similar values of eosinophils correspond to markedly different treatment effect estimates. In this example a patient with a screening eosinophil count of 499 cells/μL and another with 500 cells/μL are estimated to achieve a 18% and 33% reduction in exacerbations, respectively, when there is a negligible difference between the two eosinophil values.

The relationship between exacerbation rate reduction with mepolizumab and screening eosinophil count has been analysed using fractional polynomial modelling and the results are shown in Fig. 15.3. Here the best fitting model was an FP2

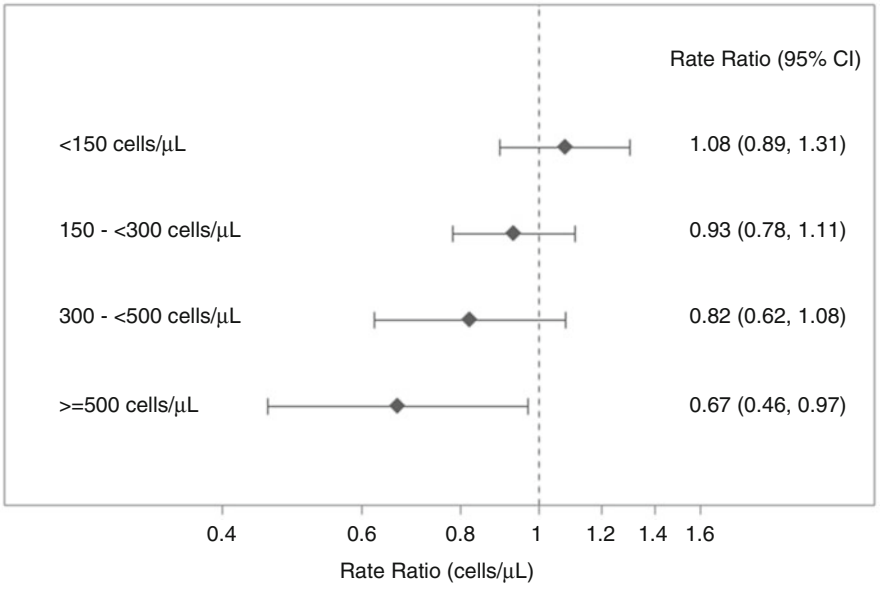

Fig. 15.2 Rate of moderate/severe exacerbations by screening blood eosinophil count: METREO/METREX trials

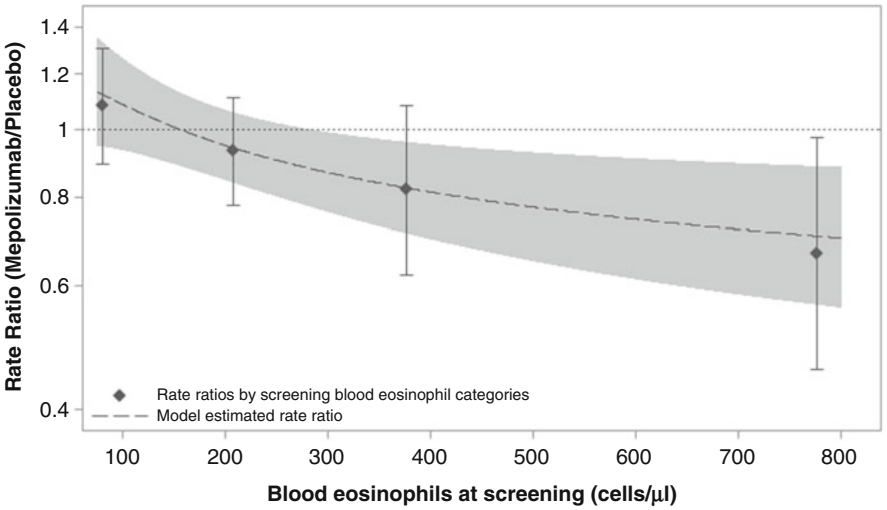

Fig. 15.3 Fractional polynomial modelling of exacerbations by screening eosinophils

function with repeated powers of (−0.5, −0.5). Estimates from the analysis based on the categories in Fig. 15.2 are overlaid on the FP plot against the mean eosinophil level in each category. The most obvious difference is that FP modelling estimates a smooth treatment effect curve across the range of eosinophils rather than a biologically implausible step-function obtained from a subgroup analysis, thus allowing more accurate estimates of treatment efficacy to be made at specific eosinophil values.

15.4.5 Splines

An alternative method to model treatment interactions with a continuous covariate is using splines. With splines and unlike FPs, the covariate is subdivided at cut-points defined as 'knots' and then separate regression curves are modelled in each segment using polynomial functions. These piecewise polynomials are anchored at the knots in such a way that the resulting curve is smooth and continuous. Various approaches are available for spline modelling but one of the more common methods is restricted cubic splines (Durrleman and Simon 1989). With this approach polynomial functions are fitted in each segment. A cubic function is used as this is the smallest degree polynomial which allows an inflection. Since cubic splines are likely to behave poorly at the tails due to lack of data, the splines are 'restricted' to be linear outside the two boundary knots. This can give an advantage over FPs, which as mentioned above can behave erratically in the lower tail particularly if values of the covariate are close to 0. Similarly, since functions are estimated in intervals of the covariate, splines may be less prone to outliers of the covariate compared to FPs.

An obvious additional step for splines is the need to specify the number and location of the knots, much like categorization in subgroup analysis. The choice of the number of knots can depend on the sample size and the prior belief in how 'undulating' the relationship is between efficacy and the covariate. Too many knots can lead to overfitting while too few can impede the flexibility of the modelling and thus might miss a true non-linear association. Authors have suggested using between 3 and 5 knots depending on sample size (Harrell 2001; Croxford 2016). For the location of the knots, Harrell (2001) has suggested particular quantiles depending on the number of knots to ensure that there is sufficient data within each interval to estimate the cubic function. For instance, for three knots Harrell recommends using the 10th, 50th, and 90th percentiles of the covariate, while for five knots use the 5th, 27.5th, 50th, 72.5th and 95th percentiles.

Despite the above guidance, the choice of knots can affect the resulting curve and so restricted-cubic splines can suffer from similar issues to subgroup analysis of the covariate. It is therefore important to pre-specify the knots where possible. Alternatively, penalized splines use many knots but discourage overfitting by restricting model complexity based on some penalty parameter (Eilers and Marx 1996). For instance, one option is to choose the spline which minimises the AIC (Binder et al. 2013). Penalised splines therefore avoid the need to specify the number and location of the knots, and hence some of the potential pitfalls of restricted cubic splines.

With these approaches, unlike FP modelling, there is currently no suitable procedure for simultaneously selecting functional forms and variables in a multivariable procedure (Binder et al. 2013). Binder et al. (2013) in their comparison of splines and FPs, concluded that for large sample sizes, the two methods often estimated similar curves, while for moderate sample sizes, FPs tended to outperform splines and were easier to implement.

Restricted cubic spline models were applied to the mepolizumab trial described above to model the efficacy of mepolizumab versus placebo on exacerbations by screening blood eosinophil count. Figure 15.4 shows the resulting curves for a spline with three knots and another with five knots using the percentiles as suggested by Harrell (2001) above and compares these curves to the best fitting FP2 function presented in Fig. 15.3. The 3-knot spline resulted in estimates curve very close the FP2 function while the 5-knot spline was more variable, likely due to fewer data points between knots. This demonstrates the need to carefully pre-specify the number of knots up-front, as the estimated curve can be sensitive to this choice.

15.4.6 Shrinkage Methods

As discussed above, when there is no true difference in efficacy between subgroups, spurious interactions can arise. This is especially the case if many subgroups are assessed, or if a specific subgroup contains a large number of levels. Subgroups including a small amount of data are particularly susceptible to showing a difference

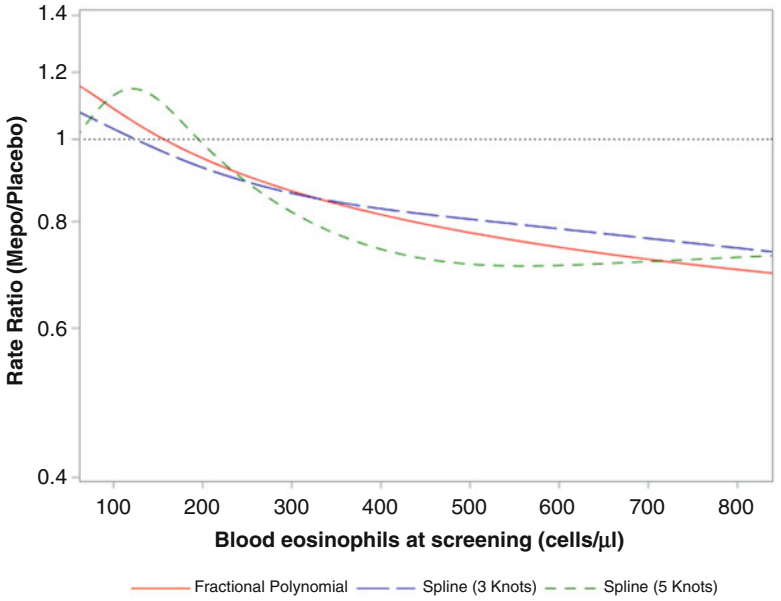

Fig. 15.4 Comparison of modelling of exacerbations by screening eosinophils using fractional polynomials with two powers and splines with three and five knots

to the complementary group due to the higher variability. Although the estimate in any one subgroup does not have a statistical bias in isolation, focusing on the specific result for that subgroup ignores relevant information from other groups.

Shrinkage methods are a technique to incorporate this information and move subgroup estimates toward the overall effect. They also increase the precision of the estimates by borrowing information across subgroups. Various shrinkage methods are available, including Empirical Bayes and Bayesian Hierarchical modelling. In the Empirical Bayes approach (Quan et al. 2013) the treatment effect d_i is first estimated in each subgroup i using data in that subgroup only. The subgroup estimates are then combined in a random-effects meta-analysis to obtain an estimate of the overall treatment effect, d, and the level of heterogeneity between the subgroup estimates as measured by the between-subgroup variability, τ^2. Subgroup estimates are then moved toward d by taking a weighted average of the original estimate d_i and d with weights w_i and $(1 - w_i)$ respectively where $w_i = \tau^2/(\tau^2 + s_i{}^2)$ and $s_i{}^2$ is the estimated variance of within-subgroup effect d_i. The result is that the original subgroup estimates are shrunk towards the overall effect, with this shrinkage being larger the higher the variability between estimates.

Another approach, Bayesian hierarchical modelling (Spiegelhalter et al. 1999), assumes that the effect in each subgroup d_i is a random quantity drawn from some common distribution centred around the overall treatment effect, d, i.e. $d_i \sim \mathrm{N}(d, \tau^2)$. The subgroup effects are assumed to be exchangeable in that there is no reason a priori to believe that the effect in one group will be different from another. Prior

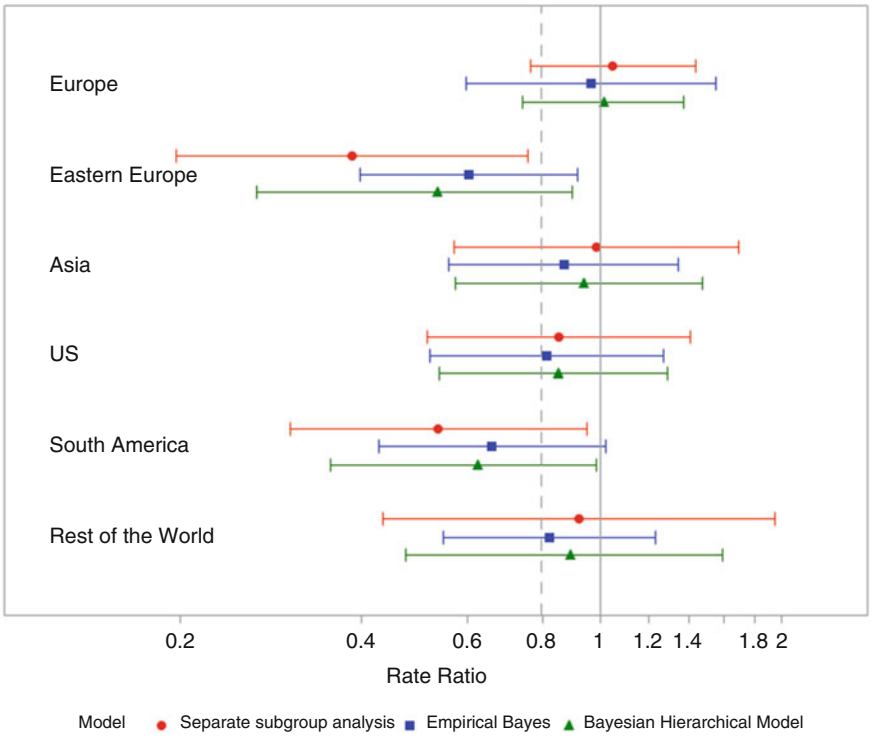

Fig. 15.5 Rate of moderate/severe exacerbations by region: METREO trial

distributions are placed on the random effect and the hyperparameters d and τ^2 to then estimate posterior distributions for the d_i and corresponding credible intervals to provide shrunken estimates of the subgroup effects.

Figure 15.5 shows a subgroup analysis of exacerbation rates by region for one of the example trials (METREO) described above. In the standard analysis using data within each subgroup separately, there appears to be a more beneficial effect of treatment in the Eastern Europe region compared to other regions, and the effect looks more favourable than the overall rate ratio of ~0.80 but confidence intervals are wide. The Empirical Bayes estimates are somewhat closer to the overall effect and the confidence intervals of most estimates are also narrower due to the borrowing of additional information from other regions. The Bayesian hierarchical analysis estimates are slightly closer to those from the original analysis, and CIs also have similar width. Thus, shrinkage techniques can incorporate prior scepticism about observing large positive or negative effects in subgroups which are unlikely to be true.

The above approach is useful primarily for evaluating a specific covariate as each patient needs to be included in a single category i.e. subgroups must be disjoint. If there is interest in assessing multiple subgroups simultaneously then patients need

to be split by the covariates of interest (e.g. male European smokers). This is likely to lead to groups containing few patients, thus affecting the stability of the model. Instead, approaches involving model averaging of subgroup-specific models can be used (Bornkamp et al. 2017). Subgroups are assessed in individual models and the model averaging applies shrinkage across all groups.

15.4.7 Bayesian Dynamic Borrowing

One novel technique which may become increasingly useful in evaluation of subgroup effects is Bayesian dynamic borrowing (Schmidli et al. 2014; Gamalo-Siebers et al. 2017). As described above, it is often required to show evidence of effect in a subgroup alongside an overall positive effect. A separate analysis of the subgroup in question is limited by sample size and does not take account of the information on the effects of treatment in the complementary subgroup. A Bayesian statistical approach is one natural quantitative method to explicitly borrow information from the complementary subgroup to provide inferences on the subgroup under evaluation.

The approach works as follows. A robust mixture prior is constructed as a weighted combination of an informative prior and a non-informative prior. The results from the complementary subgroup are used for the informative prior for the response in the subgroup of interest. The non-informative prior consists of a weak prior distribution centred on a mean of zero, reflecting no relevance of results from the complementary subgroup. This weighted combination of priors allows for dynamic borrowing of prior information; the analysis learns how much of the complementary subgroup prior information to borrow based on the consistency between the subgroup of interest and the complementary subgroup.

The prior weight, w, assigned to the informative prior component represents the prior degree of confidence in the similarity of the two subgroups. At lower prior weights the mixture prior presents a heavier tailed distribution with more prior weight being applied to the non-informative weak prior component. When the mixture prior is combined with the observed efficacy data, w is updated using Bayes theorem according to how consistent the data in the subgroup are with the complementary subgroup; the stronger the evidence of consistency, the greater the increase in the posterior weight (w^*) relative to the prior weight (w). Conversely, when there is prior-data conflict, w^* will be lower than w and will tend to zero as evidence of conflict increases, so that the informative prior is down-weighted and posterior inference is based almost entirely on the observed data in the subgroup.

To assess the strength of prior belief in the consistency assumption required to show efficacy in the subgroup, a tipping point analysis can be carried out to identify how much prior weight (w) needs to be placed on the complimentary subgroup component of the robust mixture prior for the estimate of efficacy in the subgroup of interest to show statistically significant evidence of treatment benefit (in a Bayesian

framework, this corresponds to a posterior probability that there is a treatment benefit of greater than 97.5%).

Subgroups that may be suitable for use of this dynamic borrowing approach include those subgroups of specific regulatory interest e.g. sex, race, region. For example, in a trial which includes both paediatric and adult subjects, there may be insufficient paediatric subjects to show statistical significance if this subgroup is analysed separately. A Bayesian dynamic borrowing approach of the adult data would allow assessment of the degree of belief needed that adult efficacy applied to paediatrics in order to conclude that there was evidence of efficacy in the paediatric subgroup.

15.4.8 Partitioning Methods

When there are more than a few pre-defined covariates, e.g. when there are multiple biomarkers under consideration, selection methods based on stepwise regression approaches become increasingly problematic. If there is interest in investigating complex models which go beyond evaluating relationships between treatment and a single covariate then stepwise regression may not be feasible due to the substantial number of potential two-way and three-way covariate interactions (Ruberg and Shen 2015). If there is more than one continuous covariate under evaluation, then a cut-point approach may be needed for the additional continuous variable and this brings the disadvantages described above.

Cluster analysis approaches group patients rather than examine covariates in series. They aim to identify subgroups of patients whose responses are more similar (in some sense) to each other than to those in other groups and the output is a classification tree. Historically cluster analysis has sometimes been performed with the aim of finding subgroups where the p-value for the difference between treatments is maximised, but such approaches have poor reproducibility. A more promising method is the SIDES (Subgroup Identification based on Differential Effect Search) method described by Lipkovich et al. (2011) and by Lipkovich and Dmitrienko (2014).

SIDES is a recursive partitioning method to establish response to treatment in patient subpopulations. The idea is to build a collection of subgroups by recursively partitioning a database into two subgroups at each parent group, such that the treatment effect within one of the two subgroups is maximised compared with the other subgroup. The process of data splitting continues until a predefined stopping condition has been satisfied.

An alternative approach to identify subgroups of patients with enhanced benefit is the virtual twins method described by Foster et al. (2011). The procedure works by first building a model to predict the response on treatment and control for each patient. Each patient comprises a set of ‘twins’ who differ only by the treatment they receive. This can be done by applying a random forest to each treatment group and then using the forest for a patient’s opposite treatment to predict their response

on that treatment (Foster et al. 2011). Random forests are particularly useful for this step as they exhibit low bias and prediction variance while avoiding overfitting, despite potentially dealing with a large number of covariates (Lipkovich et al. 2017).

The predicted within-patient treatment differences are then taken as observed values and used as the outcome for the subsequent subgroup identification step which uses a regression tree (or classification tree if the differences are dichotomised) to find a small number of strongly associated covariates. These are used to identify a subgroup of patients with a predicted treatment contrast greater than some clinically relevant threshold. For instance, if an asthma trial estimated the effect of treatment on FEV_1 to be 50 mL which might not be clinically relevant in many cases, then the procedure could be used to identify a subgroup likely to achieve a value more worthwhile, such as 100 mL. The enhanced treatment effect is then estimated as the difference between the effect in the subgroup and the overall population effect. Since the naïve estimate of this will be over-optimistic because the subgroup was estimated from the same data, Foster et al. (2011) describe a bias-corrected bootstrap procedure to obtain a better estimate of the effect.

Concerns can arise that clustering algorithms such as SIDES and the virtual twins method may over-fit the available data. In order to mitigate these concerns, a common practice is to divide the data into independent training and validation datasets. It is important to ensure that the training and test data sets are balanced with respect to the treatment variable and all prespecified categorical covariates (Lipkovich et al. 2011). A treatment effect identified based on the training set is considered to be confirmed if the effect is demonstrated in the validation data set.

These methods may require large sample sizes and/or large enhanced treatment effects to identify and confirm subgroups (Foster et al. 2011). If sample size is limited, it may not be practical to divide the dataset into training and validation datasets with a separate trial required to confirm findings.

A key disadvantage of the SIDES and Virtual Twins approach is the partitioning of continuous variables above and below a specific cut-point. As described above, this implies a cliff-edge effect at the cut-point which is biologically implausible.

Machine learning approaches combine different classification trees into ensembles of trees. There is no simple output showing how patients are classified; rather multiple trees are pooled in various combinations. These methods are primarily directed at prediction of response using a large number of input variables rather than at scientific understanding of which specific baseline characteristics predict response.

15.5 Discussion

In the case of a confirmatory trial for regulatory purposes, it could be argued that the burden of proof to establish an effect in each heterogeneous subgroup is with the trial sponsor. In particular, examination of results by sex and race is increasingly emphasised e.g. there are calls for efficacy to be established separately for both

women and men. Many diseases are more prevalent in one sex rather than another; for example, trials of severe asthma have recruited a majority of female participants while COPD trials reflect the historically greater incidence of smoking among men. Depending on where trials are conducted, there is likely to be imbalance in the numbers of patients across the potential classifications of race and there can be confounding of race and region which may make it difficult to disentangle medical practice from race. Small numbers of patients in a specific race category leads to large variability as reflected in wide confidence intervals for the observed effect. Going forward there is likely to be an increasing need for recruitment to trials to reflect a greater diversity in the groups studied even if this does not reflect the relative prevalence of the disease being studied and to have larger sample sizes to allow appropriate assessment of effects in subgroups defined by sex and race. One novel approach that may be helpful is Bayesian dynamic borrowing which quantifies the degree of belief needed from the complementary subgroup to confirm efficacy in the subgroup under evaluation.

Exploratory subgroup analyses are a major scientific and statistical challenge (Peto 2011) and because of multiplicity issues it is hard to identify true quantitative interactions. Subgroup analysis should depend on the heterogeneity of the population and there should be fewer requirements for these analyses when the overall population is targeted (Keene and Garrett 2014).

Formal methods for defining consistency of effect are problematic. Tests of interaction are of limited value as they do not formally provide evidence for a lack of effect, although more emphasis could be placed on estimation of the interaction effect to direct a more rational approach to assessing consistency. Methods of subgroup analysis which strongly control type I error may be able to conclude a lack of evidence for differential effects but may not identify potentially clinical relevant differences in treatment effect. Bayesian shrinkage estimates can be helpful in the interpretation of differential subgroup effects as they balance the overall effect with that observed in the particular subgroup.

A modelling approach can be enlightening in identifying covariates which predict both the absolute level of outcome and the extent to which the treatment effect is modified in that subgroup. Newer methods such as the SIDES method allow consideration of multiple covariates and the interrelationships of these covariates on treatment effect. However, for continuous variables these methods employ a partitioning (cut-point approach).

Fractional polynomial modelling and splines allow a broad range of relationships between a continuous baseline characteristic and outcome and can show treatment interactions in greater clarity compared with categorisation of the covariate. These models of outcomes against a specific covariate avoid imposition of arbitrary cut-offs for continuous variables and can determine cut-offs for treatment based on the clinical relevance of the treatment effect observed. Thus, a modelling analysis is arguably more aligned to a stratified medicine paradigm where a specific expected treatment effect can be estimated more accurately for an individual based on their value for the covariate. Prediction intervals for an individual patient will nonetheless be wide as models summarise results of a trial over a range of values.

The key issue in subgroup analysis is whether heterogeneity can reasonably be assumed. When designing a clinical trial, it is usual to assume that a common effect size holds for all patient groups. If there is a scientific rationale for heterogeneous effects across subgroups defined by a specific characteristic, then it may be necessary to show effects of treatment separately in each subgroup which implies large increases in sample size for trials. Grouin et al. (2005) for example states: "If substantial heterogeneity of the treatment effect across subgroups is suspected at the design stage, then the whole basis of the trial is undermined."

The conundrum of subgroup analysis is therefore that consistency of effect has to be assumed at some level. The trial population is already a subgroup of possible patients who could be treated. Within that trial population, subgroups can be defined based on a specific characteristic. Analysis of this specific subgroup represents a combined effect across all other characteristics. Analysis of subgroups of subgroups is possible in theory, but in practice sample size quickly becomes very small.

In conclusion therefore, the desire for individualised medicine is never likely to be completely satisfied by examination of clinical trial data which by its nature only recruits a limited number of individuals. In general, only broad statements regarding effects of individual characteristics is likely to be possible.

References

Altman DG, Royston P (2006) The cost of dichotomising continuous variables. BMJ 332(7549):1080

Ballarini NM, Rosenkranz GK, Jaki T, König F, Posch M (2018) Subgroup identification in clinical trials via the predicted individual treatment effect. PLoS One 13(10):e0205971. https://doi.org/10.1371/journal.pone.0205971

Binder H, Sauerbrei W, Royston P (2013) Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. Stat Med 32(13):2262–2277

Bornkamp B, Ohlssen D, Magnusson BP, Schmidli H (2017) Model averaging for treatment effect estimation in subgroups. Pharm Stat 16(2):133–142

Croxford R (2016) Restricted cubic spline regression: a brief introduction. In: SAS proceedings. Institute for Clinical Evaluative Sciences, Toronto, p 5621

Dane A, Spencer A, Rosenkranz G, Lipkovich I, Parke T, on behalf of the PSI/EFSPI Working Group on Subgroup Analysis (2019) Subgroup analysis and interpretation for phase 3 confirmatory trials: white paper of the EFSPI/PSI working group on subgroup analysis. Pharm Stat 18:126–139. https://doi.org/10.1002/pst.1919

Durrleman S, Simon R (1989) Flexible regression models with cubic splines. Stat Med 8(5):551–561

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11(2):89–121

European Medicines Agency (EMA) (2002) Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99. Available at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003640.pdf. Accessed Mar 2019

European Medicines Agency (EMA) (2019) Guideline on the investigation of subgroups in confirmatory clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf. Accessed Mar 2019

FDA (October 2015) Integrated summary of effectiveness, guidance for industry. Available at https://www.fda.gov/downloads/drugs/guidances/ucm079803.pdf. Accessed Mar 2019

FDA (March 2019) Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products, guidance for industry. Available at https://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181. Accessed Mar 2019

Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. Stat Med 30(24):2867–2880

Gamalo-Siebers M, Savic J, Basu C et al (2017) Statistical modeling for Bayesian extrapolation of adult clinical trial information in pediatric drug evaluation. Pharm Stat 10(1002):1807

Grouin JM, Coste M, Lewis J (2005) Subgroup analyses in randomized clinical trials: statistical and regulatory issues. J Biopharm Stat 15(5):869–882

Harrell FE (2001) Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis. Springer-Verlag New York, Inc., New York

Hemmings R (2014) An overview of statistical and regulatory issues in the planning, analysis, and interpretation of subgroup analyses in confirmatory clinical trials. J Biopharm Stat 24(1):4–18

Hemmings R, Koch A (2019) Commentary on: subgroup analysis and interpretation for phase 3 confirmatory trials: white paper of the EFSPI/PSI working group on subgroup analysis by Dane, Spencer, Rosenkranz, Lipkovich, and Parke. Pharm Stat 18:140–144. https://doi.org/10.1002/pst.1935144

International Conference on Harmonisation (ICH) (1999) Statistical principles for clinical trials. Stat Med 18:1905–1942

International Conference on Harmonisation (ICH) (November 2017) E17: General principles for planning and design of multi-regional clinical trials. https://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17EWG_Step4_2017_1116.pdf. Accessed Mar 2019

IQWiG (July 2017). General methods, version 5. https://www.iqwig.de/en/methods/methods-paper.3020.html. Accessed Mar 2019

ISIS-2 (1988) Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (second international study of infarct survival) collaborative group. Lancet 2(8607):349–360

Keene ON (1995) The log transformation is special. Stat Med 14(8):811–819

Keene ON, Garrett AD (2014) Subgroups: time to go back to basic statistical principles? J Biopharm Stat 24(1):58–71

Keene ON, Jones MRK, Lane PW, Anderson J (2007) Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the TRISTAN study. Pharm Stat 6:89–97

Li Z, Chuang-Stein C, Hoseyni C (2007) The probability of observing negative subgroup results when the treatment effect is positive and homogeneous across all subgroups. Drug Inf J 41(1):47–56

Lipkovich I, Dmitrienko A (2014) Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J Biopharm Stat 24(1):130–153

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search (SIDES): a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 30(21):2601–2621

Lipkovich I, Dmitrienko A, B D'Agostino R Sr (2017) Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 36(1):136–196

Pavord ID, Chanez P, Criner GJ, Kerstjens HA, Korn S, Lugogo N, Martinot JB, Sagara H, Albers FC, Bradford ES, Harris SS, Mayer B, Rubin DB, Yancey SW, Sciruba FC (2017) Mepolizumab for eosinophilic chronic obstructive pulmonary disease. N Engl J Med 377(17):1613–1629

Peto R (2011) Current misconception 3: that subgroup-specific trial mortality results often provide a good basis for individualising patient care. Br J Cancer 104(7):1057–1058

Pocock SJ, Assmann SE, Enos LE, Kasten LE (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 21(19):2917–2930

Quan H, Li M, Shih WJ, Ouyang SP, Chen J, Zhang J, Zhao PL (2013) Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. Stat Med 32(10):1691–1706

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat 43:429–467

Royston P, Sauerbrei W (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 23(16):2509–2525

Royston P, Sauerbrei W (2007) Improving the robustness of fractional polynomial models by preliminary covariate transformation: a pragmatic approach. Comput Stat Data Anal 51(9):4240–4253

Royston P, Sauerbrei W (2009) Multivariable model-building. Wiley, Hoboken

Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25(1):127–141

Ruberg SJ, Shen L (2015) Personalized medicine: four perspectives of tailored medicine. Stat Biopharm Res 7(3):214–229

Ruof J, Dintsios CM, Schwartz FW (2014) Questioning patient subgroups for benefit assessment: challenging the German Gemeinsamer Bundesausschuss approach. Value Health 17(4):307–309

Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B (2014) Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 70(4):1023–1032

Song Y, Chi GY (2007) A method for testing a prespecified subgroup in clinical trials. Stat Med 26:3535–3549

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR (1999) An introduction to Bayesian methods in health technology assessment. BMJ 319:508–512

Sun X, Briel M, Walter SD, Guyatt GH (2010) Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 340:850–854

Thomas M, Bornkamp B (2017) Comparing approaches to treatment effect estimation for subgroups in clinical trials. Stat Biopharm Res 9(2):160–171

Yusuf S, Wittes J, Probstfield J, Tyroler HA (1991) Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 266(1):93–98