Article:
Dane, A. orcid.org/0000-0002-2997-4074, Spencer, A. orcid.org/0000-0002-4194-3561, Rosenkranz, G. et al. (3 more authors) (2019) Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the EFSPI/PSI working group on subgroup analysis. Pharmaceutical Statistics, 18 (2). pp. 126-139. ISSN 1539-1604

https://doi.org/10.1002/pst.1919

Pharmaceutical Statistics

Subgroup Analysis and Interpretation for Phase 3 Confirmatory Trials: White paper of the EFSPI/PSI Working Group on Subgroup Analysis

Journal:	*Pharmaceutical Statistics*
Manuscript ID	PST-16-0105.R4
Wiley - Manuscript type:	Main Paper
Date Submitted by the Author:	24-Sep-2018
Complete List of Authors:	Dane, Aaron; DaneStat Consulting Limited Spencer, Amy; Statistical Services Unit, University of Sheffield Rosenkranz, Gerd; Medizinische Universitat Wien Lipkovich, Ilya; Quintiles Innovation, Parke, Tom; Berry Consultants LLC
Key Words:	subgroup analysis, regulatory labelling, bias adjustment, late-phase clinical programs
Abstract:	Subgroup by treatment interaction assessments are routinely performed when analysing clinical trials and are particularly important for Phase 3 trials where the results may affect regulatory labelling. Interpretation of such interactions is particularly difficult, as on one hand the subgroup finding can be due to chance, but equally such analyses are known to have a low chance of detecting differential treatment effects across subgroup levels, so may overlook important differences in therapeutic efficacy. EMA have therefore issued draft guidance on the use of subgroup analyses in this setting. Although this guidance provided clear proposals on the importance of pre-specification of likely subgroup effects and how to use this when interpreting trial results, it is less clear which analysis methods would be reasonable, and how to interpret apparent subgroup effects in terms of whether further evaluation or action is necessary. A PSI/EFSPI Working Group has therefore been investigating a focused set of analysis approaches to assess treatment effect heterogeneity across subgroups in confirmatory clinical trials which take account of the number of subgroups explored, and also investigating the ability of each method to detect such subgroup heterogeneity. This evaluation has shown that the plotting of standardised effects, bias-adjusted bootstrapping method and SIDES method all perform more favourably than traditional approaches such as investigating all subgroup-by-treatment interactions individually or applying a global test of interaction. Therefore, these approaches should be considered to aid interpretation and provide context for observed results from subgroup analyses conducted for Phase 3 clinical trials.

Subgroup Analysis and Interpretation for Phase 3 Confirmatory Trials: White paper of the EFSPI/PSI Working Group on Subgroup Analysis

Authors: Aaron Dane[1], Amy Spencer[2], Gerd Rosenkranz[3], Ilya Lipkovich[4] and Tom Parke[5] on behalf of the PSI/EFSPI Working Group on Subgroup Analysis

[1]DaneStat Consulting, Macclesfield, UK, [2]Statistical Services Unit, University of Sheffield, UK, [3]Medical University of Vienna,[4]Quintiles, USA, [5]Berry Consultants, UK.

Corresponding author: Aaron Dane, Email: aarondane@danestat.com, +44 7821 720 631

Running head: PSI WG evaluation of Subgroup Analysis approaches

Abstract

Subgroup by treatment interaction assessments are routinely performed when analysing clinical trials and are particularly important for Phase 3 trials where the results may affect regulatory labelling. Interpretation of such interactions is particularly difficult, as on one hand the subgroup finding can be due to chance, but equally such analyses are known to have a low chance of detecting differential treatment effects across subgroup levels, so may overlook important differences in therapeutic efficacy. EMA have therefore issued draft guidance on the use of subgroup analyses in this setting. Although this guidance provided clear proposals on the importance of pre-specification of likely subgroup effects and how to use this when interpreting trial results, it is less clear which analysis methods would be reasonable, and how to interpret apparent subgroup effects in terms of whether further evaluation or action is necessary.

A PSI/EFSPI Working Group has therefore been investigating a focused set of analysis approaches to assess treatment effect heterogeneity across subgroups in confirmatory clinical trials which take account of the number of subgroups explored, and also investigating the ability of each method to detect such subgroup heterogeneity. This evaluation has shown that the plotting of standardised effects, bias-adjusted bootstrapping method and SIDES method all perform more favourably than traditional approaches such as investigating all subgroup-by-treatment interactions individually or applying a global test of interaction. Therefore, these approaches should be considered to aid interpretation and provide context for observed results from subgroup analyses conducted for Phase 3 clinical trials.

Key words: Subgroup analysis, regulatory labelling, bias adjustment, late-phase clinical programs

Introduction

It is common to conduct subgroup analyses in clinical trials during all phases of drug development, and to try to understand heterogeneity in treatment effect across various levels of subgroup factors of interest. This is of relevance in early phase studies in order to better understand which patients to study in future trials, whilst for late stage clinical trials is applied in order to understand the effects of a given treatment across a range of baseline factors. Consistency of treatment effect across trial subgroups indicates that the conclusions made regarding treatment benefit are applicable across various baseline characteristics and associated subpopulations, whilst substantial heterogeneity in treatment effect may suggest clinically relevant differential treatment effects across subpopulations. At the extreme, the presence of substantial heterogeneity may imply that the conclusion of beneficial treatment effect is only relevant for a subset of the population.

Interpretation of subgroup analyses is difficult [1] as any apparent heterogeneity can be due to chance, which is particularly likely when a large number of subgroup analyses are undertaken. Conversely, clinical trials are generally not designed for detecting subgroup heterogeneity, so statistical tests may miss important interactions due to low power [2]. This is particularly challenging as approaches need to account for both issues simultaneously, as illustrated by Gonnermann [3] who also concluded that further work was needed to understand alternative statistical approaches in this area. Despite these issues, potential subgroup differences cannot be ignored, so it is still necessary to understand potential subgroup heterogeneity [4], and as such, consideration needs to be given for how such effects are to be analysed and interpreted.

Given these issues, EMA proposed draft guidance in February 2014 [5] and held a Workshop in November 2014 to address these considerations. This guidance was generally well received, but a key area of further work related to providing information on methodological approaches and criteria to conclude consistency of effect. Two European Pharmaceutical industry statistical organizations, the European Federation for Statisticians in the Pharmaceutical Industry (EFSPI) and Statisticians in the Pharmaceutical Industry (PSI) therefore formed a working group to further explore possible analysis approaches in order to

inform subgroup evaluation in a regulatory environment. As this work is prompted by the EMA draft guidance and possible regulatory labelling, approaches have focused on assessing heterogeneity of treatment effect across subgroup levels in a Phase 3 trial. As such, this paper assumes the situation where the analysis of the overall population showed a positive effect before any subgroup effects were explored.

This paper summarises the key activities of the working group, and their possible implications for future regulatory evaluation. As the focus is to understand whether heterogeneity of treatment effect across subgroup levels is real, any reference to subgroup analysis or subgroup effects will be in relation to this heterogeneity. This paper discusses predictive effects only (factors driving a differential effect across treatments) and does not consider prognostic factors (factors predictive of outcome, regardless of treatment). However, factors that are known to be prognostic for the disease of interest would be a more plausible set of candidate variables to explore for treatment effect heterogeneity.

Pre-specification of subgroup effects

When conducting subgroup analyses it is often necessary to assess the heterogeneity of treatment effect across various levels of a number of subgroup factors. When conducting late stage clinical trials there are a variety of reasons for this, ranging from a standard regulatory authority requirement, their importance in a particular clinical setting, or to address key reimbursement questions. Indeed, FDA requires that subgroups defined by gender, age and race are to be analysed and dosage modifications to be identified for specific subgroups [6], and in 2014 the agency issued an action plan "to enhance the collection and availability of demographic subgroup data" [7]. Given the multiplicity challenges associated with interpreting a large number of subgroup analyses, the first objective for regulators and sponsors alike is to attempt to minimize the number of subgroups analysed wherever possible, and regulators at both EMA and FDA have recommended that the focus should be on the subgroups of primary interest [5, 8].

All subgroups to be analysed in this way must be pre-specified, and it will often still be necessary to choose a fairly large number of subgroups within a confirmatory clinical trial. Therefore, it is also important to provide a clear justification for the plausibility of a

particular subgroup by treatment interaction. This will ease interpretation and will also make for a stronger argument in relation to any subgroup effect found. Any prior evaluation of plausibility should be based on historical data, external trial data obtained from literature or the labels of approved products, or data from earlier trials in the clinical program (Phase I or II) for the project under study. Further, it is important to specify the nature of the interaction (direction of impact on overall outcome), and to state which changes are clinically relevant and would impact labelling. From such pre-specification, it is possible to consider the subgroups in 3 categories. The following categories below are similar to those outlined in the EMA guidance:

Confirmatory (strong reason to expect a significant, clinically relevant heterogeneous response):
In this case it is necessary to address the differential levels of effect for a specific factor explicitly and is addressed in the design and interpretation of the trial (for example, through type I error adjustment). This category will not be covered further in this paper.

Biologically plausible (existing biological rationale or external evidence of heterogeneous response):
In this case, a heterogeneous response would be a differential treatment effect across subgroups levels considered clinically meaningful, and as such would impact product labelling. There are likely very few subgroups in this category, and EMA guidance states that it is relevant to discuss and plan for an assessment of consistency of effects. This could well be a situation where subgroups may have some prognostic effect, but there has been no evidence of a treatment by subgroup interaction to date.

Hypothesis generating only (no prior evidence to expect a heterogeneous response):
This would include all other subgroups necessary for study based upon standard regulatory requirements, or clinical practice. As these subgroups are not anticipated to show any differential treatment effect, approaches to understand the effects in the context of the large number of subgroups evaluated can be considered.

This categorization of subgroup analyses is helpful both when planning subgroup analyses and when interpreting the results. For a subgroup to have some plausibility there must be a

prior hypothesis regarding a potential treatment-by-subgroup interaction. In particular, the "biologically plausible" category represents hypothesized subgroup factors or baseline covariates which may govern differential treatment response. These factors may be simple (e.g., based on a single baseline characteristic) or complex (i.e., based on multiple baseline characteristics simultaneously.) The number of proposed factors in this category is expected to be small and identification of these factors, along with specific hypotheses regarding direction and magnitude of anticipated differences are key to evaluating the robustness of any subgroup findings once the results are available.

Current approaches

At present, it is common to conduct a statistical test for treatment-by-subgroup interaction by adding these terms to the primary analysis model. This can involve assessing subgroup by treatment interactions for a large number of subgroups, which inflates the chances of a false positive finding [9].

Further, interaction tests have low power, a point raised within the EMA guidance which states that lack of statistical significance of an interaction is not sufficient to conclude subgroup by treatment homogeneity. An approach to improve the power when testing for an interaction is to use arbitrary criteria such as $p<0.10$. Whilst this makes it easier to detect an interaction, this also increases the chances of false positive findings. Indeed, if 10 independent subgroup factors are explored in a trial, none of which have an actual effect, the probability of observing at least one significant interaction at $p<0.10$ is $1 - (1 - 0.10)^{10}$, or approximately 65%. It is acknowledged that it could equally be argued that a more stringent type I error rate should be applied as a result of the multiple testing of many subgroups [10]. It is for this reason that the simulation exercise described has explored the type I error rate, defined as the number of times a subgroup by treatment interaction is incorrectly identified under the "null" that no such interaction is present.

An alternative approach is to perform variable selection in regression models, where parameters associated with various interaction effects are added or removed from the model depending upon variable importance. However, interpretation of this approach can be challenging as very different conclusions can be reached dependent upon the selection

method applied (e.g., backward or forward selection). Furthermore, variable selection is unstable where small perturbations in the data may lead to selection of different variables. Inference has often been done without consideration of model selection, i.e., as if the selected model had been pre-specified. Only recently exact methods have been developed for type I error control or confidence intervals for parameter estimates after selection [11]. These methods are still under debate. Therefore, this approach has not been explored further.

One way of providing some type I error control when investigating multiple subgroups is to perform a global test of interaction (comparing model fit of models with no interaction terms, and all interaction terms of interest), and this is justified when there is no biological rationale to expect an interaction for the subgroups studied.

It is also common practice to produce univariate forest plots for each subgroup to help understand the effect of an individual subgroup on overall response, and forest plots including a reference line indicating the point estimate for the overall effect [12] will provide information on any subgroups whose 95% CI excludes this point estimate for the overall treatment effect. When interpreting such forest plots, those subgroup levels with a 95% CI looking very different from the overall result would often require further discussion and investigation of whether the overall result applies equally to all levels of this subgroup. EMA guidance specifically addresses the use of forest plots but suggests caution as a formal rule for interpretation that is "both sensitive enough to detect heterogeneity …and specific [enough to detect true subgroups] is not available". Further, as such plots do not include information on confounding of key covariates, they should generally be interpreted alongside other methods to understand the likelihood of such findings in the presence of other important factors.

Alternative methods considered

There is a wealth of literature describing possible approaches to subgroup analysis. Lipkovich et al [13] have recently provided a review of the literature, much of which is focused on the exploratory setting. Similarly, given the importance of subgroup analysis, the whole of issue 24 of the Journal of Biopharmaceutical Statistics in 2014 was dedicated to subgroup analysis in clinical trials [14] and prominent authors have provided a tutorial on the

statistical considerations of such analyses [15]. There have also been other papers which have described ideas related to the use of permutation distributions in other settings, such as that to assess regression to the mean [16,17], which although not directly related do show the approach has a wide applicability for generating "null" distributions in the context of data driven subgroup evaluation.

Given the purpose of this paper has been to identify approaches that could be used in the setting of a confirmatory clinical trial and for regulatory labelling, approaches were explored that were consistent with current practice, but also addressed the concerns regarding the power to detect true subgroup heterogeneity, whilst controlling the rate of incorrect identification of a subgroup due to multiplicity. As such, we consider a new method which produces a plot of standardised effects and will provide context for the magnitude of subgroup heterogeneity seen in relation to what would be expected by chance. This approach is motivated by the desire to find an approach akin to a forest plot, but which takes account of the number of subgroups analysed and the correlation between these subgroups. In addition, we present approaches that either control type I error or provide estimates of treatment effect after adjusting for subgroup selection, both of which are important when providing information to regulatory agencies, treating physicians or for reimbursement.

This section will outline these approaches, whilst the following sections will describe a simulation study used to evaluate the operating characteristics of these methods in different scenarios.

(a) Plot of Standardised effects

A resampling based graphical method to present Standardised Effects Adjusted for Multiple Overlapping Subgroups (SEAMOS) is proposed by Dane [18]. SEAMOS is intended to provide a graphical presentation of the results for all pre-specified subgroups, and to illustrate how extreme the results are expected to be by chance, given the number of subgroups analysed and the correlation between the subgroups. This approach can be used when it is possible to clearly divide a subgroup into categories, as would be expected in an evaluation used to assess pre-specified subgroups likely to impact regulatory labelling. Briefly, observed standardised effects for the difference between the subgroup level and overall effect

are ordered from largest to smallest to highlight the statistically most extreme standardised effects in either direction. These are then compared to a "null" distribution of standardised effects, or the distribution of effects expected by chance when there are no subgroup-by-treatment interactions. This is calculated by using a resampling algorithm and is achieved by randomly permuting the rows of covariates many times against the vector made up of the response variable and the treatment identifier. This has the effect of removing any subgroup by treatment effects but preserves the overall treatment effect and the correlation between subgroups. The smallest and largest standardised effects from each permutation are taken and used to produce a probability interval of extreme effects. If there were no subgroup heterogeneity, we would expect the largest standardised effect to lie within this probability interval 95% of the time. Similarly, we would expect the smallest standardised effect to lie within the interval 95% of the time. These intervals can be used to assess the highest and lowest observed values, with points lying outside these probability intervals requiring further evaluation, as shown in Figure 1, controlling the chances of incorrectly identifying a subgroup when there is no subgroup heterogeneity at ≤10%. Similarly, probability intervals can be constructed for all ordered subgroups as described in Dane [18].

Given there are often challenges in using a formal method which makes firm conclusions regarding statistical assessment criteria, the intention is to provide a method which places the results into context in addition to providing a basis for further exploration of subgroups. There are some approaches available that take a similar approach, but only work when the subgroup levels are all fully independent (for example, when exploring the effects by country in the clinical trial) [19,20,21,22].

(b) Adjusted subgroup estimates via bootstrapping

This bootstrapping method proposed by Rosenkranz [23,24] is primarily based on estimation. The method fits a reference model containing treatment and all factors as main effects. For each factor defining a subgroup, the method fits a model containing in addition the treatment by factor interaction effect and picks the factor/model leading to the best fit to the data. As it is acknowledged that the estimates provided for subgroup are exaggerated, particularly when a large number of subgroups are investigated, this selection process is repeated on a series of bootstrap samples of the original data to account and adjust for such bias, with the degree of adjustment being dependent upon the amount of times the factor in question is selected from

the bootstrap samples (for example, if a factor is picked in every bootstrap sample very little adjustment would occur, whilst much more adjustment would occur if the factor were selected less frequently). The bootstrap samples are then used to calculate an interaction effect estimator (and its standard deviation) adjusted for selection bias and model uncertainty.

The method allows for all data types and in principle also for continuous variables. Instead of selecting only the factor with the best fit in most bootstrap samples, all factors selected more often than expected by chance could be selected. Furthermore, the full population (i.e., the model containing no treatment by factor interaction term) is part of the selection process such that all subgroups compete with the full population in terms of goodness-of-fit, meaning it is possible to select the model with no treatment effect heterogeneity.

(c) SIDES

The SIDES method (Subgroup Identification based on Differential Effect Search) [25,26] applies the following simple search strategy combined with a resampling approach in order to adjust any significance level (or p-value) for the model selection approach. This method was originally developed for controlling type I error in a more exploratory setting, (with subgroups possibly defined as "signatures" of up to 3 continuous variables with data-driven cut-offs) but could equally be applied (as a special case) to the setting where the search is restricted to subgroups based on a single binary factor selected from a candidate set. This method can also be used for all response types and is applied as follows:

- All candidate subgroups are evaluated and the covariate with the larger differential effect across subgroup levels is selected. The subgroup level requiring further evaluation is that which has the larger treatment effect.
- Then the null reference distribution for the treatment effect p-value of the selected subgroup is constructed by randomly permuting the treatment labels and applying the above procedure to the resulting data.
- This is repeated 1000 times and the adjusted p-value is determined as the proportion of reference sets where the selected subgroup's p-value is less than or equal to that found in the observed data.
- For the purposes of this evaluation, a subgroup is selected if the adjusted p-value is below the pre-specified cut-off of 0.1. Thus, the procedure ensures that the type I error of the entire selection strategy is within 10%.

- To make the null distributions of p-values insensitive to the overall treatment effect the analysis data set is standardised before applying the procedure. This is achieved by subtracting the respective treatment arm's mean and dividing by the arm-specific standard deviation.

Simulation exercise to evaluate performance of the methods

Simulation scenarios used to evaluate operating characteristics

When investigating the performance of various methods, two broad simulation scenarios have been used, the first ("simple set") where all subgroup factors were simulated independently and the second ("complex set") where some correlation higher than that expected by chance was generated. For both approaches the operating characteristics will focus on the situation with a total of 10 binary subgroup factors. Note that we have not considered scenarios where the number of subgroup factors is much less than this (2 or 3 factors, for example), as this is not felt to be a realistic scenario for Phase 3 confirmatory trials where it is necessary to address a number of regulatory and clinical questions regarding subgroups. Similarly, we have not considered scenarios of an interaction between two subgroups and with treatment, as we aimed to simulate a situation where a Phase 3 trial is to be conducted and no subgroup effects are anticipated, so such a scenario seems unlikely.

Given all of these approaches have the potential to identify more than one subgroup as part of the procedure this may involve identifying both a correct and an incorrect subgroup. For the purposes of operating characteristic calculation, a second step is performed which explores which subgroup has the most compelling result - this "primary subgroup" is used to define whether a correct (or incorrect) subgroup identification has occurred. The only exception is for SEAMOS. As this approach presents information on both levels of each subgroup factor there is a small chance that two different subgroup factors will be identified as the most extreme positive and most extreme negative effects. In this case they have both exhibited results beyond those expected by chance and would both need to be explored further. As such, it is possible to identify both a correct and incorrect subgroup in this case.

The methods used for the simulation scenarios are similar to those described with full details in Dane [18]. Briefly, this involved performing 1,000 simulations for each scenario and assessing how often a true subgroup was identified (in scenarios with true treatment effect heterogeneity for one subgroup factor), or how often a subgroup was incorrectly identified when there was no treatment effect heterogeneity across subgroups. This number of simulations was chosen as a balance between time taken for the simulation exercise and achieving sufficient precision, and is considered reasonable as the standard deviation around a 10% error rate and 1,000 simulations is estimated to be $0.95\% = \sqrt{\frac{(0.1*0.9)}{1000}}$. This was felt sufficient for ensuring the error rates were reasonably well controlled.

The simulations used assumptions regarding an overall treatment effect and an enhanced effect for one level of the subgroup (S+ group), defined as k, where k took possible values of 1-, 2- or 4-times the overall treatment effect). Individual patient outcomes were simulated from a normal distribution using the mean based on the allocation to treatment arm and to subgroup *S*+ or *S*-. In the "simple set" of simulations, for all other subgroups, defined by the other 9 factors, the overall effect, θ, is assumed. Using these assumptions, 4 different sized trials were considered by varying the magnitude of effect and size of study, and within each, separate simulations were performed according to the proportion of patients in the subgroups. Subjects had either a 0.1, 0.25 or 0.5 probability of being in the *S*+ subgroup based on the first factor, and this same proportional split was used for all other factors. This magnitude of effect was assessed under "high power" and "normal power" scenarios, where "normal power" is defined as 90% power to detect $p<0.05$ (2-sided) for the overall effect, and "high power" as 90% power to detect an effect with $p<0.01$. Further, each method was assessed when a smaller or larger trial was required for these scenarios to understand whether the ability to detect heterogeneity was affected by having a smaller number of patients within the subgroup. This resulted in sample sizes of 200 and 270 patients per group for smaller sample size and the normal and high-power scenarios respectively, and 760 and 1080 patients per group for the larger sample sizes.

The previous text describes the simple simulations, in which the 10 subgroup factors were all generated independently. The "complex set" of simulations were produced which introduced some inter-relationship between the true factor and one correlated "noise" factor. In this case

the correlation of these two binary variables was calibrated in order to achieve 50% overlap (See details in the Appendix).

Results for the "simple" and "complex" set of simulation scenarios were very similar. Therefore, the following section presents plots of the operating characteristics for only the complex scenario as we feel this is likely to be closer to real data observed in a clinical trial. Similar plots using a "subject level" evaluation which estimates how well the identified subgroup from the data captures the patients belonging to the "ideal" subgroup (see Appendix for further details) gave similar conclusions to those described in this paper.

When considering these scenarios, it is also important to consider the size of treatment effect in the remaining population ($\boldsymbol{\theta}_{S-}$), and hence whether the interaction is qualitative or quantitative. This is important as it can have implications for regulatory labelling and will depend upon the proportion of subjects in the subgroup positive group (P_{S+}) and the magnitude of differential effect in this positive subgroup level (k). Some examples are presented when the overall treatment effect is 5 units (see Table 1). As can be seen in Table 1, when one subgroup level has a very large positive effect (e.g., k=4), the effect in the remainder of the population may be either zero (for P_{S+}=0.25) or negative (for P_{S+}=0.5). This must be considered alongside a review of the performance of the analysis methods, as there are key implications regarding benefit-risk and whether a new treatment should be approved in all patients.

Results of simulation exercise

In order to understand the performance of the methods outlined in this paper it is necessary to assess how often a given approach correctly identifies subgroup heterogeneity and how often a subgroup is incorrectly identified when there is no heterogeneity. Using the simulation scenarios described, operating characteristics were calculated for traditional interaction testing, the global interaction test and the methods outlined in the "Alternative methods" section. Because traditional interaction testing is the only one of these methods which does not attempt to correct for multiple testing, the results for this method are presented separately in the "Performance of traditional interaction testing" section, whilst the evaluation of the

remaining methods are included in the section summarising "Performance of methods which account for the evaluation of multiple subgroups".

Performance of traditional interaction testing

The performance of the more traditional approach of fitting a multivariate model including all subgroups involved defining a subgroup of interest if any of the subgroup-by-treatment interaction terms from the statistical model were statistically significant using the criteria $p<0.1$. Table 2, below demonstrates that, as expected, when 10 subgroups are explored and criteria of $p<0.1$ is applied the type I error is very high. Table 2 also demonstrates that even when there is true heterogeneity within one subgroup, other subgroups without this heterogeneity are incorrectly identified on many occasions. This characteristic is concerning when no subgroups interactions are anticipated, particularly when decisions are required regarding labelling for such subgroups. The performance characteristics for the alternative methods are presented in the next section and show that these methods perform more favourably than the approach of assessing all interactions individually in this way.

Performance of methods accounting for the evaluation of multiple subgroups

The operating characteristics for the scenarios with 10 subgroup factors, where one of the "unimportant" factors has an overlap of 50% with the one for which the "true" effect was generated are given in Figure 2. The panels a-d show the results for four sample size and overall effect size combinations.

The evaluation of the various approaches for the scenarios with 10 subgroup factors, both with independent subgroups (not shown) and with an enhanced correlation between the true subgroup and one other subgroup showed that the global interaction (GI) test controlled the type I error at approximately 10%. SEAMOS also controlled the type I error at 10% for subgroup sizes 0.1 and 0.25, and at a slightly lower rate for a subgroup size of 0.5. The bootstrapping method tended to result in higher type I error with the smaller sample sizes of 200 or 270 patients, but this higher type I error was not seen with larger sample sizes. Meanwhile, SIDES showed lower type I error than the other approaches across all scenarios explored.

For each method, the power to detect true subgroup heterogeneity increased with true subgroup size (P_{S+}), subgroup effect (k) and sample size (for the same true effect size), as would be expected. Comparing the power of the different methods, for the scenarios where the true treatment effect in one subgroup level was twice as large as the treatment effect for the population as a whole (k=2), the bootstrapping method and then SEAMOS had the highest power, whilst SIDES had lower power along with the lower type I error stated previously. However, SIDES performs more preferably than the GI test at the lower subgroups sizes of 0.1 and 0.25. When the true treatment effect within the better performing level of the subgroup was 4-times the treatment effect in the population as a whole (k=4), all three of the newer methods had similar power to detect the subgroup by treatment interaction and performed better than the GI test.

When many subgroups are analysed simultaneously and there is a subgroup with a truly heterogeneous effect it is also possible to incorrectly identify another subgroup as exhibiting treatment effect heterogeneity. As a result, we also explored the number of times an incorrect subgroup was identified within the simulation scenarios when there was true subgroup heterogeneity for one subgroup factor. The results from this exploration were similar to those presented previously, in that the SIDES method tended to give the lowest rates of incorrect identification and SEAMOS also tended to give lower values than the GI test. Regarding the bootstrapping method, the degree of incorrect identification again tended to depend on the parameters of the simulation scenario. In the scenario with subgroup effects twice that of the overall effect, where the overall effect size was 0.33 and subgroup size was 0.1 or 0.25, this method gives comparatively high incorrect identification rates, whilst for other simulations the results are similar to SEAMOS.

Conclusions and discussion

The EMA draft guidance is a great advancement in terms of the pre-specification and interpretation of subgroup analyses. The work adopted by this Working Group is looking to add guidance on how to best analyse the data when a reasonably large number of subgroups are considered, and when the results have the potential to impact regulatory labelling and patient access to medicines.

With any subgroup analysis, pre-specification of the more plausible subgroup-by-treatment interactions is key to effective interpretation. This is a point EMA have stated very clearly in their guidance [5], and a paper by FDA authors [8] also talks of reducing the number of subgroups. Our paper considers the situation where effects are, a priori, assumed to be consistent (or sufficiently consistent to apply the overall result to that subgroup), and this premise would ideally be agreed with regulatory agencies at the design stage of a trial. By this we mean that the subgroups to be explored can be pre-specified in the "Biologically plausible" or "Hypothesis generating only" categories. This investigation would be performed separately within these two categories. Situations where there is a stronger belief regarding subgroup heterogeneity have not been considered, as this should lead to a different study design, and not be addressed at the analysis stage.

When exploration of the "Biologically plausible" or "Hypothesis generating only" categories is required, methods such as SEAMOS, SIDES and the adjusted effect size estimates via bootstrapping provide useful context when interpreting subgroup analyses and compare favourably with traditional approaches which test all subgroups independently and have very high error rates.

The graphical SEAMOS method and bias-adjustment methods presented should be considered key tools to help quantify whether subgroup findings are real. They have benefits over traditional methods of analysis such as assessing all interactions separately, stepwise regression or a global test of interaction, as they appropriately account for the number of subgroups analysed and the correlation between subgroups. It should be noted that we do not advocate the use of any of these methods in isolation, but rather suggest presenting their results along with the observed effects to provide context when assessing the likelihood that any subgroup heterogeneity is real.

The situation when each of these three methods will be most useful will depend upon the key aims for the specific analysis in question. For example, the number of subgroups being explored, or the degree of focus on type I error control will affect when each method should be adopted. Similarly, whether the subgroups are categorical or continuous (or whether there is an established method of categorizing continuous various) will be critical. This is particularly true of SEAMOS which, although providing very useful context for the observed

results, can only be undertaken when a clear method of dividing a subgroup into categories is possible. However, such categorization is most likely in the case of the late phase confirmatory clinical trials considered here. Finally, the most appropriate method will also depend upon whether the aim of additional subgroup work is to provide context for the observed results (SEAMOS), provide adjusted estimates of treatment effect (bias-adjusted bootstrapped estimates) or provide adjusted p-values of subgroup effects (SIDES).

The simulation scenarios investigated have included a set of independent subgroups and those with a degree of overlap above that expected by chance. It is acknowledged that there are many other possible scenarios with respect to the overlap (or correlation) between subgroups, and it has not been possible to assess all of these. However, we have no reason to believe we would see differences with regard to the conclusions regarding type I error and power. Additional simulation scenarios could be the subject of further work, but it would also be important for anybody applying these approaches to demonstrate the type I error control based upon observed subgroups and the relationship between them.

The approach to error control in this setting is a balance between controlling the rate of incorrect subgroup identification (or incorrect labelling restrictions), with incorrectly allowing a broad label. The approaches explored have looked to control the incorrect subgroup identification error at ~10%. Other values could be used in discussion with regulatory agencies, and would depend upon the magnitude of overall effect, the therapeutic index and number of other therapeutic options in that particular situation.

Bayesian methods have been considered extensively when evaluating subgroups [27,28,29,30,31,32] and are appealing as they incorporate prior beliefs regarding subgroup heterogeneity into the evaluation. In our preliminary work we considered the "Simple Regression" and "Dixon and Simon" methods reviewed in Jones [28] as these methods appeared promising in the setting we have outlined in this paper. We encountered issues related to model fitting and did not to pursue this further at that time, but additional work is ongoing exploring a range of Bayesian approaches to provide recommendations on how such methods may be applied to the regulatory setting with pre-defined subgroups.

This paper has focused on the situation when only one confirmatory Phase 3 trial has been conducted, as any unexpected subgroup heterogeneity is most challenging to interpret in that

setting. When a number of studies have been conducted, this will provide additional, critical information in terms of whether the heterogeneity of treatment effect is replicated, or whether the magnitude of differential treatment effect is consistent across trials. When heterogeneity is not replicated this casts more doubt on the likelihood of a differential effect and should be used alongside an assessment of biological plausibility when interpreting whether the heterogeneity is more likely to be real, or a chance finding.

These methods could also be used for post-hoc evaluation of efficacy, or when the overall treatment effect is not positive. This special case has not been addressed in this paper as this work is in the context of the regulatory setting where an overall treatment effect would be necessary before investigating consistency in subgroups. Further, post-hoc subgroup analysis and/or analysis in the setting of a negative treatment effect can only be considered hypothesis generating as the implications for type I error are less clear, and the lack of pre-specification of likely subgroup effects (and hence plausibility) makes such results very difficult to interpret.

In summary, the approaches investigated in this paper regarding SEAMOS, SIDES and the adjusted effect size estimates via bootstrapping all provide useful context when investigating and interpreting subgroup heterogeneity and, when presented alongside the observed data and used in conjunction with a clear pre-specification of more plausible subgroups, are useful tools for interpreting apparent subgroup effects. Ongoing work is defining how these approaches can be used within the regulatory review setting, whether they can contribute in the review of reimbursement dossiers, and how such statistical approaches can be used to inform an assessment of the benefit risk profile of a new treatment.

Acknowledgments.

The conclusions described within this document represent the work of the European Federation for Statisticians on the Pharmaceutical Industry (EFSPI) and Statisticians in the Pharmaceutical Industry (PSI) Working Group on Subgroup analysis. The work of the Project Team is gratefully acknowledged, especially because it depends upon the generous contributions of personal time. The EFSP/PSI Working Group members are as follows: Aaron Dane (WG lead, DaneStat Consulting), Chrissie Fletcher (Amgen), Heiko Goette (Merck), Necdet Gunsoy (GSK), Ilya Lipkovich (Quintiles), Henrik Loft (Lundbeck), Brian

Millen (Lily), Tom Parke (Berry Consultants), Arne Ring (Medac), Gerd Rosenkranz (Medical University of Vienna), Amy Spencer (University of Sheffield), David Svensson (AstraZeneca).

The initial ideas for this work were developed when Aaron Dane and Amy Spencer were employed by AstraZeneca, and Gerd Rosenkranz was employed by Novartis and under funding provided by the UK Medical Research Council, Project No. MR/M005755/1. Therefore, we would like to thank AstraZeneca and Novartis for their support in this research.

References

[1] Wittes, J. On looking at subgroups. Circulation. 2009; 119: 912-915.

[2] Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. Statistics in medicine—reporting of subgroup analyses in clinical trials. New England Journal of Medicine, 2007; 357(21), 2189-2194.

[3] Gonnermann A, Kottas M, Koch A. Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation. Bundesgesundheitsbl - Gesundheitsforschung - Gesundheitsschutz. 2015; 58: Issue 3, pp 274-282

[4] Koch, A., Framke, T. Reliably basing conclusions on subgroups of randomized clinical trials. Journal of Biopharmaceutical Statistics. 2014; 24:1, 42-57.

[5] EMA. Guideline on the investigation of subgroups in confirmatory clinical trials. European Medicines Agency/Committee for Medicinal Products for Human Use. 2014; EMA/CHMP/539146/2013.

[6] FDA, Content and format of a new drug application. 1985; 21 CFR 314.50[1]

[7] FDA, FDA action plan to enhance the collection and availability of demographic subgroup data. 2014.

[8] Alosh M, Fritsch K, Huque M, Mahjoob K, Pennello G, Rothmann M, Russek-Cohen E, Smith F, Wilson S, Yue L. Statistical Considerations on Subgroup Analysis in Clinical Trials. Statistics in Biopharmaceutical Research; 2015; 7(4): 286-303.

[9] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355(9209): 1064–69

[10] Lagakos SW. The challenge of subgroup analyses - reporting without distorting. New England Journal of Medicine 2006; 354;16: 1667-69.

[11] Tibshirani RJ, Taylor T, Lockhart R, Tibshirani R. Exact PostSelection Inference for Sequential Regression Procedures, JASA 2016, 111, 600-620.

[12] Cuzick, J. Forest plots and the interpretation of subgroups. The Lancet, 2005; 365(9467), 1308.

[13] Lipkovich I, Dmitrienko A, D'Agostino Snr R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine; 2017; 36(1); 136-196.

[14] Wang, Sue-Jane, and Alex Dmitrienko. "Guest Editors' Note: Special Issue on Subgroup Analysis in Clinical Trials." Journal of biopharmaceutical statistics. 2014; 24(1): 1-3.

[15] Alosh M, Huque MF, Bretz F, D'Agostino RB Sr. tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials. Statistics in medicine 2017 36(8): 1334-1360.

[16] Farlow MR, Small GW, Quarg P, Krause A. Efficacy of Rivastigmine in Alzheimer's Disease Patients with Rapid Disease Progression: Results of a Meta-Analysis (Subgroup analysis with corrected p-values for a regression to the mean effect). Dementia and Geriatric Cognitive Disorders, 20(2-3): 192-197, 2005.

[17] Krause A, Pinheiro J. Modeling and Simulation to adjust p-values in Presence of a Regression to the Mean Effect. The American Statistician, 61 (4): 302-307, 2007.

[18] Dane A, Spencer A, Stone A, Svensson D. The use of Standardised Effect Plots when Interpreting Analyses of Overlapping Subgroups. Submitted to Statistics in Biopharmaceutical Research, 2017.

[19] Anzures-Cabrera J. and Higgins, JPT. Graphical displays for meta-analysis: An overview with suggestions for practice. Res. Synth. Methods, 2010; 1(1): 66–80.

[20] Galbraith, Rex. Graphical display of estimates having differing standard errors. Technometrics, 1988; 30 (3): 271–281.

[21] Chen J, Zheng H, Quan H, Li G, Gallo P, Soo PO, Binkowitz B, Ting N, Tanaka Y, Luo X, Ibia E, and for the Society for Clinical Trials (SCT). Multi-regional Clinical Trial (MRCT) Consistency Working Group. Graphical assessment of consistency in treatment effect among countries in multi-regional clinical trials. Clin Trials, 2013; 10(6): 842-851.

[22] Schou I. M., and C. Marschner I. Methods for exploring treatment effect heterogeneity in subgroup analysis: an application to global clinical trials, Pharmaceut. Statist, 2015; 14(1), 44–55.

[23] Rosenkranz, GK. Bootstrap corrections of treatment effect estimates following selection. Computational Statistics and Data Analysis, 2014; 69, 220–227.

[24] Rosenkranz, GK. Exploratory subgroup analysis in clinical trials by model selection. Biometrical Journal, 2016; 58(5), 1217-1228.

[25] Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G.. Subgroup Identification based on Differential Effect Search (SIDES): A recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine, 2011; 30(21): 2601–2621.

[26] Lipkovich, I., Dmitrienko, A. Strategies for Identifying Predictive Biomarkers and Subgroups with Enhanced Treatment Effect in Clinical Trials Using SIDES, Journal of Biopharmaceutical Statistics, 2014; 24:1, 130-153.

[27] Millen BA, Dmitrienko A, Song G. Bayesian Assessment of the Influence and Interaction Conditions in Multipopulation Tailoring Clinical Trials. Journal of Biopharmaceutical Statistics, 2014; 24:1, 94-109.

[28] Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. Clinical Trials, 2011; 8(2): 129–143.

[29] Berger JO, Wang X, Shen L. A Bayesian Approach to Subgroup Identification. Journal of Biopharmaceutical Statistics, 2014; 24: 110–129.

[30] Henderson NC, Louis TA, Rosner GL, Varadhan V. Individualized Treatment Effects with Censored Data via Fully Nonparametric Bayesian Accelerated Failure Time Models, 2017. arXiv:1706.06611v1 [stat.ME], available at https://arxiv.org/abs/1706.06611

[31] Gu X, Chen N, Wei C, Liu S, Vassilike A, Herbst RS, Lee JJ. Statistics in Biosciences, 2016; 8:1; 99-128.

[32] Hsu Y-Y, Zalkikar J, Tiwari RC. Hierarchical Bayes approach for subgroup analysis. Statistical Methods in Medical Research, July 2017. https://doi.org/10.1177%2F0962280217721782

Table 1 Effect sizes in subgroup positive group and in complement of that subgroup for the effect sizes and subgroup sizes outlined in the simulation scenarios.

Overall treatment effect ($\boldsymbol{\theta}$)	Differential Effect in positive subgroup level (k)	Treatment Effect for subgroup positive subgroup level ($\boldsymbol{\theta}_{S+}$)	Size of positive subgroup level (P_{S+})	Treatment effect in remaining population ($\boldsymbol{\theta}_{S-}$)
5 units	4 x overall	20 units	0.1	3.33 units
			0.25	0 units
			0.5	-2.5 units
	2 x overall	10 units	0.1	4.44 units
			0.25	3.33 units
			0.5	0 units

Table 2. Operating characteristics for traditional analysis assessing treatment-by-factor interaction terms at p<0.10

Overall Treatment Effect/standard deviation	N per arm	Subgroup size (% of overall population)	Type I error, %	Probability of detecting true subgroup (power) [Probability of detecting incorrect subgroup], %	
				2x overall effect	4x overall effect
0.167	760	10%	64.3	20.7 [60.0]	81.1 [60.3]
		25%	67.5	45.5 [62.0]	99.9 [61.2]
		50%	65.7	90.9 [62.2]	100 [61.0]
	1080	10%	64.8	26.7 [62.9]	91.3 [60.4]
		25%	64.4	60.2 [63.3]	100 [61.4]
		50%	63.5	97.8 [60.7]	100 [59.4]
0.33	200	10%	66.4	21.9 [58.8]	76.4 [62.1]
		25%	63.2	46.3 [57.5]	99.9 [60.8]
		50%	67.5	91.7 [58.6]	100 [60.6]
	270	10%	63.0	25.8 [60.7]	90.2 [61.3]
		25%	66.4	56.7 [59.9]	100 [60.6]
		50%	63.3	97.6 [59.2]	100 [64.1]

Notes:

1. Estimates in this table are approximate, and are based upon 1000 simulations; Simulations present number of significant treatment x covariate interaction effects ($p < 0.10$) in traditional regression analysis based on 10 covariates (1 factor with "real" treatment interaction and 9 factors with no interaction).
2. The probability of at least 1 of 9 factors with no interaction showing a significant interaction effect is $1 - ((1 - 0.10)^9) \approx 61\%$ (similar to squared brackets in columns "2x overall effect" and "4x overall effect").
3. The probability of at least 1 of 10 factors with no interaction showing a significant interaction effect is $1 - ((1 - 0.10)^{10}) \approx 65\%$ (similar to column "Type I error, %").

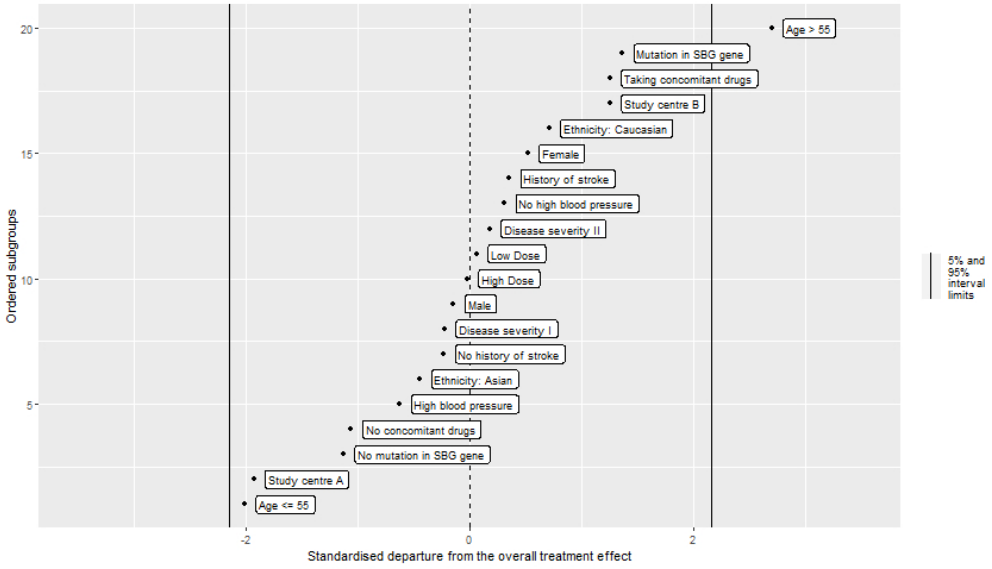

Figure 1: Example SEAMOS plot with probability intervals for the most extreme effects.

296x169mm (72 x 72 DPI)

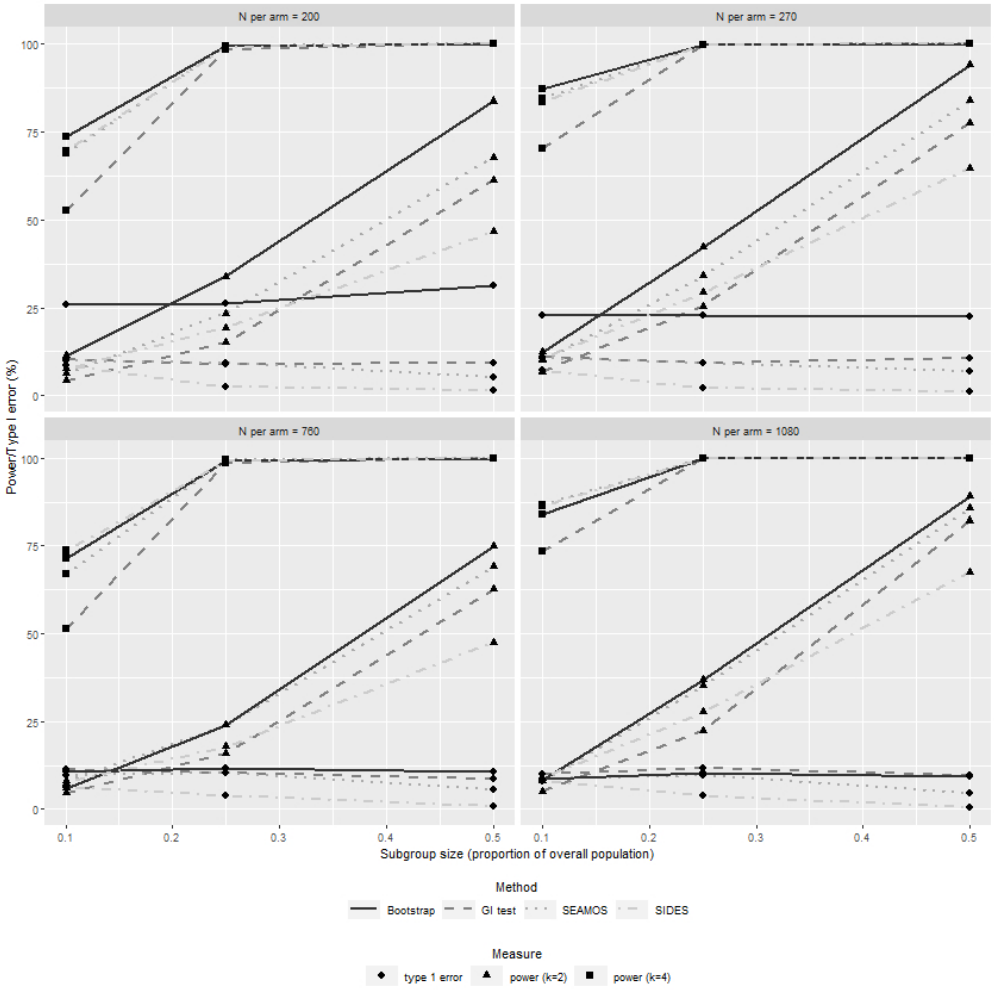

Figure 2: Summary of operating characteristics comparing the subgroup analysis approaches under investigation.
Panels present four sets of simulation results for sample sizes of 200 per group and 270 per group, both with an overall large treatment size; and 720 per group and 1080 per group, both with an overall small treatment size.

296x296mm (72 x 72 DPI)

APPENDIX

Framework for simulation scenarios when simulating binary subgroups with desired overlap

Summary

Following the simple simulation scenarios with 10 subgroup factors simulated independently, further work regarding more complex scenarios was explored. Here we describe the methods used for simulating these slightly more complex scenarios. The basic set-up is as follows:

- A number of subgroup factors is again 10, two of which explain a degree of the subgroup by treatment interaction.
- Simulate data assuming one of these subgroups has an important interaction as before.
- Simulate the second subgroup on the basis of the correlation/overlap between subgroups 1 and 2. This assumes an overlap of 0.5, which is higher than expected by chance for all the simulation scenarios.
- All other subgroups are simulated independently of these first two, as previously.
- Reapply for all previous scenarios in terms of sample size, subgroup size, overall treatment effect and subgroup treatment effect.
- Define performance criteria on identifying either the "correct" subgroup (1) or the "correct" patients (those in subgroup 1).

Simulation of correlated variables

Let $X_1,..,X_{10}$ be binary covariates defining "elementary" subgroups. We define the "overlap" measure between two binary covariates X_1 and X_2 using Jaccard coefficient as the ratio of the probability that a subject belongs to both subgroups to the probability that s/he belongs to either of them.

$$J_{12} = \frac{\Pr(\{X_1 = 1\} \cap \{X_2 = 1\})}{\Pr(\{X_1 = 1\} \cup \{X_2 = 1\})}$$

The amount of overlap is a measure of correlation between binary variables, and can be calculated using the underlying multivariate normal distribution.
The correlation can be easily evaluated numerically for the case of bivariate normal.

Criteria of subgroup identification performance evaluation

It is possible to consider two criteria that measure performance of the different analysis methods at:

- Subgroup level. How well was the predictive subgroup identified?
- Subject level. How closely are the subjects who are be in the "true" subgroup approximated with the identified subgroup?

The methods presented in this paper have focussed on the subgroup level identification rate. The subject level identification rate shows higher success rates for all methods, but does not change the conclusions regarding the relative performance of the methods.

Subgroup level

Here we evaluate how well a subgroup identification method can identify true predictors. This is essentially the same analysis used in the simple scenario, but considers which (if any) of the subgroups are picked out by the analysis and uses the following measures:

- Type I error: In the case where k=1 (no subgroup effect), if a subgroup is identified in the analysis, this is a type I error.
- Power: Where there is a subgroup effect (k>1), if the correct subgroup (S1) is identified, this contributes to power.
- Incorrect ID: Where there is a subgroup effect, if any subgroup other than the correct one (not S1) is identified, this is an incorrect ID.

Subject level

Here we evaluate how close the identified subgroup is to some ideal one in the sense that the identified subgroup captures most of the patients defined by the ideal subgroup (even if it may fail to capture the correct subgroup factor). We use standard measures that are used in evaluating the quality of predicting a target binary outcome (here the true subgroup, S) by the identified subgroup, $\hat{S}$. In this case, we define the true subgroup S as subgroup 1, so this method will give excellent results when subgroup 1 is identified, good results when subgroup 2 (the highly correlated subgroup) is identified and poor results when another subgroup (based on one of the markers uncorrelated with the true subgroup) is identified. The performance measures we use are:

- sensitivity= $\left|\hat{S} \cap S\right| / |S|$
- specificity = $\left|\hat{S}^c \cap S^c\right| / \left|S^c\right|$, where S^c is complement of S.