WILEY Statistics in Medicine

A recursive partitioning approach for subgroup identification in individual patient data meta-analysis

Dipesh Mistry [iD] | Nigel Stallard | Martin Underwood

Warwick Medical School, University of Warwick, Coventry, UK

Correspondence
Dr Dipesh Mistry, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK.
Email: d.mistry@warwick.ac.uk

Background: Motivated by the setting of clinical trials in low back pain, this work investigated statistical methods to identify patient subgroups for which there is a large treatment effect (treatment by subgroup interaction). Statistical tests for interaction are often underpowered. Individual patient data (IPD) meta-analyses provide a framework with improved statistical power to investigate subgroups. However, conventional approaches to subgroup analyses applied in both a single trial setting and an IPD setting have a number of issues, one of them being that factors used to define subgroups are investigated one at a time. As individuals have multiple characteristics that may be related to response to treatment, alternative exploratory statistical methods are required.

Methods: Tree-based methods are a promising alternative that systematically searches the covariate space to identify subgroups defined by multiple characteristics. A tree method in particular, SIDES, is described and extended for application in an IPD meta-analyses setting by incorporating fixed-effects and random-effects models to account for between-trial variation. The performance of the proposed extension was assessed using simulation studies. The proposed method was then applied to an IPD low back pain dataset.

Results: The simulation studies found that the extended IPD-SIDES method performed well in detecting subgroups especially in the presence of large between-trial variation. The IPD-SIDES method identified subgroups with enhanced treatment effect when applied to the low back pain data.

Conclusions: This work proposes an exploratory statistical approach for subgroup analyses applicable in any research discipline where subgroup analyses in an IPD meta-analysis setting are of interest.

KEYWORDS
individual patient data, meta-analysis, randomised controlled trial, subgroup analysis

1 | INTRODUCTION

Randomised controlled trials typically evaluate the performance of an intervention with the aim of answering a primary hypothesis, or research question, on the effect of treatment. Secondary subgroup analyses may subsequently be

performed to identify subgroups that most, or least, benefit from treatment. For instance, in the area of drug development, subgroup analyses of predictive biomarkers have gained much popularity in recent years and are considered an important part of the drug development process. Identifying subgroups in this way can help inform decision making thus improving individualized patient care by targeting treatment accordingly.

Statistical issues associated with subgroup analyses, in particular, lack of power and multiplicity are well known.[1] It is therefore important to adhere to recommendations and proposed subgroup analysis guidelines to ensure that analyses are of a credible standard.[2-7] However, conventional approaches to subgroup analyses have some limitations. First, the most common statistical approach for performing subgroup analyses is to fit a regression model including a treatment-covariate interaction term that tests participant characteristics one at a time. However, in reality, trial participants have multiple characteristics that also need to be investigated either simultaneously or in a systematic stepwise fashion. Secondly, it is recommended that a clear distinction is made between confirmatory subgroup analyses and exploratory subgroup analyses. Confirmatory analyses investigate a small number of well-defined pre-specified subgroups thus limiting multiplicity, whereas exploratory analyses investigate a larger number of loosely defined subgroups. Part of defining subgroups in both these analyses involves selecting cut-points for continuous and categorical covariates that are to be pre-specified and clearly justified at the outset based on current clinical knowledge. However, limiting the number of subgroups explored and selecting cut-points solely based on clinical knowledge could result in important alternative subgroups defined by other covariates and cut-points going unnoticed. The aforementioned issues therefore clearly suggest that alternative and more sophisticated exploratory statistical approaches are required to identify all potential subgroups. The potential subgroups, provided they are clinically or biologically plausible, can then be tested using standard methodology before being accepted.

In general, there are a broad range of statistical methods that have been developed for conducting exploratory and confirmatory subgroup analysis.[8,9] In the field of data mining, a number of data-driven approaches exist that offer a different approach for performing subgroup analyses. In particular, tree-based methods are popular approaches used when the aim is to identify subgroups with high and low outcome. Many of these methods are based on the classification and regression tree (CART) methodology for identifying predictors of outcome in cohort studies proposed by Breiman et al.[10] More recently, other advanced variants of the CART type approach, also referred to as subgroup discovery methods, have been developed that specifically perform subgroup analyses in a single trial setting. Key methods include the following: Virtual Twins, STIMA, QUINT, GUIDE, Interaction Trees, and SIDES.[11-16] A recent tutorial nicely describes, compares, and summarises the key features of these various methods.[17] This paper will focus on one these methods; the Subgroup Identification based on Different Effect Search (SIDES) method.[16]

One could power a trial for subgroup analyses; however, such a trial might be unfeasibly large. An alternative approach would be to pool together and analyse individual patient data (IPD) from several similar trials. IPD meta-analyses provide an ideal framework with improved statistical power to investigate and identify subgroups.[18] The SIDES method, however, has only been developed and implemented in a single trial setting. It cannot be applied directly to an IPD meta-analysis setting as the method would ignore the trial level clustering inherent in the data. In an IPD setting with a hierarchical data structure, tree methods that utilise fixed or mixed effects modelling are required. A number of tree-based methods have been proposed for data with a correlated structure,[19-24] but these methods are only applicable in a longitudinal or repeated measures setting where data are collected over time. Hajjem et al[25] proposed a mixed-effects regression tree for hierarchical or nested data; however, this approach identifies trees that best predict response rather than identify subgroups with differential responses to treatment. It is more clinically useful to identify meaningful subgroups than to predict response. This paper therefore proposes an extension to the SIDES method, as a novel exploratory approach for subgroup identification in an IPD meta-analyses framework. The extended method will be referred to as the IPD-SIDES method.

The structure of this paper is as follows. Section 2 describes the example that motivated the proposed methodological extension in this paper. Section 3 begins with a general introduction to the concept of recursive partitioning and tree-based methodology followed by a description of the original SIDES method. Section 4 will then detail the proposed extension of the SIDES method (IPD-SIDES) for application to IPD meta-analyses data. Thereafter, the results of simulation studies are presented in Section 5 to demonstrate how the proposed method performs in an IPD meta-analysis setting. Section 6 provides an example of the proposed method applied to real clinical trial data. Finally, in Section 7, some concluding remarks will be provided.

2 │ MOTIVATING EXAMPLE

The methodological development in this paper was motivated by the challenge of finding subgroups of patients who most benefit from the available recommended therapist delivered interventions for non-specific low back pain

(NSLBP).[26-29] This work was part of a project that collected a repository of IPD from 19 existing NSLBP trials testing similar interventions to then go on to perform subgroup analyses. Because subgroup analyses in individual NSLBP trials are generally underpowered and of a poor quality,[30] a large repository of similar data was collected to provide improved statistical power to better undertake the task at hand.

The aim of the subgroup analyses using the proposed tree-based method is to identify subgroups of patients with the greatest treatment benefit in terms of their back-related disability. If such subgroups do exist, then this will aid decision making and matching treatments to those patients who present with NSLBP who are most likely to benefit. Targeting treatment in this manner would mean increased effectiveness of the treatment when compared with the average.

This paper will look at a pooled dataset consisting of data on 4540 individuals from 4 trials to illustrate how the approach might work when applied to a real dataset. There were 3 demographic and baseline covariates that were common across the 4 trials: age, gender, and baseline quality of life. The quality of life was measured using the SF-36 questionnaire which was also recorded at short-term follow-up for all 4 trials. The SF-36 questionnaire measures health-related quality of life and when scored comprises of 2 aggregated summary measures, namely, the mental component score (MCS) and physical component score (PCS).[31] The MCS and PCS scores range from 0 to 100 where a lower score represents poorer mental or physical functioning. The proposed methods will be applied to these data later on where the change from baseline to short-term SF-36 MCS and SF-36 PCS scores will be analysed as 2 separate dependent variables with age, gender, and the baseline value of the dependent variable used as the set of baseline covariates.

3 | EXISTING METHODOLOGY

In general, the tree growing component of many tree methods relies on a technique referred to as recursive partitioning which utilises a splitting criterion to form binary splits of the covariate space in order to grow a tree-like structure. The splitting criterion is essentially used to compute a score for any given split where either the largest score or smallest score is indicative of a better split. It thus plays a key role in determining how a tree is grown. For example, a tree method that searches for subgroups using the treatment-covariate interaction effect as the splitting criterion would identify a split with the largest score as the best split, whereas a method that uses the associated interaction effect P-value as the splitting criterion would identify a split with the smallest P-value as the best split. For a more detailed insight on recursive partitioning, one can refer to a review provided by Zhang and Singer on recursive partitioning and its applications.[32] The SIDES methods will now be described. For further detail, one can refer to the original SIDES paper.[16]

3.1 | SIDES method

1. *Growing an initial tree*

We first describe the algorithm for the SIDES procedure followed by a more detailed description of the splitting criterion and the continuation criterion. The algorithm for growing the tree is as follows:

- Start at the root node consisting of the entire dataset
- Step 1 - Evaluate the splitting criterion for all possible splits of every covariate, excluding any covariates already used to define the parent node, retaining only the best split for each covariate. Order the covariates from smallest adjusted P-value to largest adjusted P-value where the adjusted P-values are computed using the Sidak-based multiplicity adjustment which adjusts for the number of splits searched for a given covariate (see below).
- Step 2 - Select the best M covariates from the ordered best splits. The value of M is specified by the user where the recommended value is 5. For each of the M splits, form the split creating 2 child nodes and retain the child node with the larger positive treatment effect, provided it satisfies the continuation criterion. The retained nodes now become parent nodes for the next iteration.
- Step 3 – Repeat steps 1 and 2 for the newly formed parent nodes
- Step 4 – Repeat steps 1 to 3 until either a pre-specified maximum number of levels (L) is reached or if no more splits can be formed, ie, the continuation criterion is not satisfied. In both cases, the previously formed parent nodes become terminal nodes.

3.1.1 | Splitting criterion

The SIDES method uses a splitting criterion that tests the difference in the treatment effect precision between 2 child nodes with the aim of identifying the subgroup or child node with the most significant treatment effect. The splitting criterion is of the form

$$p = 2 \cdot \left[1 - \Phi \left(\frac{Z_{E1} - Z_{E2}}{\sqrt{2}} \right) \right] \tag{1}$$

where Z_{E1} and Z_{E2} are the 1-sided test statistics for the treatment effect computed for the 2 subgroups, respectively, and $\Phi \left(\frac{Z_{E1} - Z_{E2}}{\sqrt{2}} \right)$ is the cumulative distribution function of the standard normal distribution.[16] For covariates with more than 2 potential cut-points, the P-value of each evaluated potential split is adjusted by applying a Sidak-based multiplicity adjustment to overcome a well-known issue associated with tree-based methods known as variable selection bias.[33-35] The multiplicity adjustment is of the form

$$1 - (1 - p_i)^{G^*}$$

where p_i is the unadjusted P-value for the i-th split obtained using[1] and G^* is the effective number of splits computed by $G^* = G^{1 - \bar{r}}$ where G is the total number of splits that a particular covariate has and $\bar{r}$ is the average correlation of all the unadjusted P-values for all splits of that covariate.[16]

3.1.2 | Continuation criterion

The SIDES method controls the tree complexity by using a continuation criterion as part of the tree growing algorithm. In step 2 of the SIDES algorithm, a child node with a large positive treatment effect is retained only if it satisfies the continuation criterion. The continuation criterion is given by

$$p_c \leq \gamma \cdot p_p$$

where p_c is the treatment effect P-value of the child node, p_p is the treatment effect P-value of the parent node, and γ is the relative improvement parameter that controls the complexity of the tree. Prior to running the method, the user must specify the maximum number of covariates or levels L that defines a subgroup, where the recommended value is 3. This means that any identified subgroups will at most be defined by L covariates; hence, the tree will have at most L levels where $L = 0$ is the starting level, ie, the entire dataset. Each level of the tree has a relative improvement parameter value that ranges from 0 to 1 where a smaller value makes the procedure more selective. The values for each level can be either user specified or optimally selected using a cross-validation procedure as described by the authors.[16] Hence, once the relative improvement parameter values are in place, a child node is only retained provided its treatment effect P-value is less than or equal to the right hand side of the continuation criterion.

- *Selecting the final candidate subgroups*

The first step of the SIDES procedure grows the tree and produces a list of candidate subgroups. Many of these subgroups may be spurious findings or an artefact of the dataset and thus need to be removed. To control for this and assess reproducibility of subgroups, the authors propose a resampling-based procedure that is applied only once at the end after the whole tree has been grown. The resampling procedure computes an adjusted treatment effect P-value for each of the identified candidate subgroups to control the overall type I error in the weak sense.[16] This procedure fixes the covariate columns and randomly scrambles the rows of the outcome and treatment variables together in order to maintain the treatment effect and correlation structure. The SIDES procedure is then applied to this resampled dataset, and the P-value of the best subgroup is recorded. This resampling procedure is repeated many times, eg, 1000 times, to form a distribution of P-values. Adjusted P-values are then computed by calculating the proportion of P-values in the distribution obtained using the resampling procedure that are less than the observed candidate subgroup P-value. Comparing the unadjusted P-value to the adjusted P-value gives a good indication as to whether the identified subgroups are spurious or not.

4 | EXTENSION OF THE SIDES METHOD TO AN IPD META-ANALYSIS SETTING (IPD-SIDES)

4.1 | A proposed new splitting criterion

The objective of the SIDES method splitting criterion is different to what we require the method to do. In particular, the method seems to evaluate the difference in precision between 2 nodes when forming a split rather than directly evaluating the differential effect, which is what we are interested in. This became more apparent when closely inspecting the subgroups identified by the method when applied to real NSLBP data. Many of the selected subgroups had a positive treatment effect that were rather similar to its disregarded counterpart subgroup. For example, one of the splits formed 2 nodes with a treatment effect of 3.4 (SE: 0.29, $n = 3464$) in 1 node and 3.5 (SE: 0.93, $n = 381$) in the second node. Computing the test statistics in[1] gives $Z_{E1} = 11.5$ and $Z_{E2} = 3.8$, which suggests there is a large differential effect between the 2 nodes despite the treatment effects being quite similar. The resultant SIDES splitting criterion was very small ($P < 0.001$) indicating a difference that is highly significant. The same differential effect evaluated using a treatment-covariate interaction term in a regression model suggested otherwise; with an estimated interaction effect of 0.1 (SE: 0.93, $n = 3845$, $P = 0.891$), ie, a non-significant difference. This therefore highlights the need for a new splitting criterion to be defined in order to meet our objective, ie, directly assess the differential treatment effect. For this reason, we propose a new splitting criterion

$$P = 2 \cdot [1 - \Phi(Z_{int})] \tag{2}$$

where Z_{int} is the 2-sided hypothesis test statistic computed for the interaction effect estimate obtained using a linear regression model.[36] The proposed criterion computes a P-value for the interaction effect where a smaller p-value is indicative of a larger interaction effect.

The second issue identified from preliminary work was that the SIDES method tended to detect spurious subgroups when large or very large single 1-way interaction effects were present.[37] The proportion of spurious subgroups detected increased as the sample size increased. This issue was investigated to try to identify the source of the problem. Due to the sample sizes being large, the treatment effect estimates have small variability hence giving a highly significant 1-sided treatment effect approximated as being zero at level 1 of the tree for the selected subgroup. Consider the example in Table 1 where the SIDES method was applied to a simulated dataset ($N = 5000$) where there is a standardized 1-way interaction of 1.5 (very large). At level 1, the correctly selected subgroup ($X_1 > 0$) has a treatment effect of 0.81 (SE: 0.04, $P = 0.00$). The splitting criterion P-values are then used to order the next best potential splits from best to worst, ie, smallest P-value to largest P-value, regardless of whether the P-value is significant or not. Hence, non-significant differential effects are considered by the method at level 2 of the tree. Again, referring to the example in Table 1, $X_2 > 0$ was considered as the next best split at level 2 with a non-significant splitting criterion value ($P = 0.425$). As the sample size of the selected subgroup is large at level 2, the method identifies the 1-sided treatment effect as being highly significant ($P = 0.00$). The continuation criterion in this situation is satisfied as the 1-sided treatment effect P-value approximation of the selected subgroup at level 2 is equal to the P-value of the selected subgroup at level 1. Hence, a spurious subgroup is detected. A solution to control this issue was to introduce a significance threshold V in Step 1 of the algorithm. It is not necessary that a strict threshold be imposed, eg, $V \leq 0.05$, but a less stringent threshold, eg, $V \leq 0.20$, would suffice to ensure the method has some flexibility to detect plausible subgroups. The SIDES method that uses the proposed splitting criterion as defined by Lagakos[2] and that also uses the significance threshold V will be referred to as the modified SIDES method.

TABLE 1 – Example of spurious subgroup detection using the SIDES method applied to simulated data ($N = 5000$) when there is a single standardized 1-way $T \times X_1$ interaction of 1.5 (very large)

		Selected Subgroup				Disregarded Subgroup					
Level	Subgroup	n1	T_1	SE (T_1)	T_1 P-Value	n2	T_2	SE (T_2)	T_2 P-Value	Differential Effect	Splitting Criterion P-Value
1	$X_1 > 0$	2500	0.81	0.04	0.00	2500	−0.76	0.04	1.00	1.57	0.000
2	$X_1 > 0$ and $X_2 > 0$	1250	0.84	0.06	0.00	1250	0.77	0.06	0.00	0.07	0.425

4.2 | Proposed extension to allow for IPD meta-analysis

Because we now want to apply the modified SIDES method to IPD from different trials, we need to extend the method so that it adjusts for the between-trial variability. This requires us to estimate the interaction effect test statistic parameter in the splitting criterion[2] having accounted for the between-trial variation. A natural extension of the modified SIDES method is to estimate the splitting criterion by fitting a fixed trial effect in the regression model to give a model of the form

$$Y_{ij} = \beta_{0i} + \beta_1 T_{ij} + \beta_2 X_{ij} + \beta_3 T_{ij} \cdot X_{ij} + \varepsilon_{ij} \tag{3}$$

where β_{0i} is the intercept term for the i-th study, the ij subscript denotes the j-th observation in the i-th study, and ε_{ij} is a normally distributed error term with mean zero. Another option would be to fit a random effects model of the form

$$Y_{ij} = \beta_{0i} + \beta_1 T_{ij} + \beta_2 X_{ij} + \beta_3 T_{ij} \cdot X_{ij} + \varepsilon_{ij} \tag{4}$$

$$\beta_{0i} \sim N\left(\beta_0, \sigma_{\beta_0}^2\right)$$

where the study level covariate β_{0i} is set as having a random effect with mean β_0 and variance $\sigma_{\beta_0}^2$. We can thus define the splitting criterion in the same way as shown in Equation 2, where the interaction effect parameter is estimated using either a fixed-effects model or a random-effects model instead of a linear regression model. The extension of the modified SIDES method applied to IPD will be referred to as the IPD-SIDES method. The remaining components of the method, ie, complexity control and final subgroup selection, remain the same.

5 | SIMULATION STUDIES

5.1 | Simulation study design

Simulation studies were conducted to assess the performance of the proposed IPD-SIDES method compared with the modified SIDES method in detecting subgroups with enhanced treatment effect when applied to data with a hierarchical structure. The simulation studies considered both the fixed-effect model[3] and random effect model[4] for estimating the splitting criterion for the proposed IPD-SIDES method. A simple linear regression model was used to simulate data for a single trial. The model related a continuous dependent variable Y to a treatment indicator T, 5 dichotomous covariates $X_1, X_2, ... X_5$ with normally distributed error ε with $\varepsilon \sim N(0, 1)$. The overall mean for the single trial was randomly generated using a normal distribution with zero mean and variance τ^2, ie, $N(0, \tau^2)$, to incorporate between-trial heterogeneity where τ^2 is to be specified. Several single trial data were simulated separately using this model and then pooled together to form an IPD dataset. Each single trial dataset was simulated such that there was an equal proportion of observations in each quadrant of the 2×2 table for the treatment-covariate interaction. A full factorial design was used to investigate a number of different simulated scenarios by varying 4 factors; the sample size of each trial in the combined IPD dataset, the size of the interaction effect for $T \cdot X_1$, the size of the interaction effect for $T \cdot X_2$, and the between-trial variance τ^2. There were 5 trials in each pooled dataset where each trial had a fixed sample size of either 200, 500, or 1000. The proportion of individuals in each category of the treatment indicator and the covariates were assumed to be equal. For both $T \cdot X_1$ and $T \cdot X_2$, we considered standardized interaction effect sizes of 0, 0.2 (small), 0.5 (medium), and 0.8 (large). The values for τ^2 were specified as being either 0.1 (small) or 0.9 (large) as these values are inside the range of the typical between-trial heterogeneity found in IPD meta-analyses.[38,39] These values equate to an intra-cluster correlation coefficient (ICC) of approximately 0.08 and 0.42, respectively. Finally, the main effects and interaction effects for all other covariates were set as zero. All permutations in the full-factorial design were simulated 1000 times where each permutation took approximately 5 hours to run when using a fixed-effect model, and around 14 hours when using a random effect model. Simulations were performed using R software (version 3.3.2).

5.2 | Parameter specifications

Application of the methods to the simulated data required certain parameters to be pre-specified. A stopping criterion was put in place such that the minimum number of individuals in any node was 10% of the total sample. The maximum

number of levels was set to 2 ($L = 2$) as there are only 2 covariates, ie, any identified subgroups can only be defined by a maximum of 2 covariates. Moreover, the best 2 splits ($M = 2$) were considered for each node where the significance threshold was set as $V = 0.10$. The relative improvement parameters for the continuation criterion were selected using 5-fold cross validation where all permutations of the values from 0 to 1 at the first level and 0.2 to 1 at the second level were searched to find the optimum relative improvement parameter sequence. The final subgroups were selected using the re-sampling based procedure, drawing 1000 samples each time.

5.3 | Simulation results

The results of the IPD-SIDES method when using a fixed-effect model and random effect model were very similar. Hence, only the results from the fixed-effect model simulation studies will be presented. The results of the random effect model simulation studies can be found elsewhere[37] or made available upon request.

5.4 | Modified SIDES and IPD-SIDES results

The modified SIDES and IPD-SIDES simulation results for the pooled sample sizes $N = 1000$, $N = 2500$, and $N = 5000$ are presented in Tables 2, 3, and 4, respectively. The performance of both the modified SIDES and IPD-SIDES methods were found to be quite similar and within the range of variability due to simulation error in the presence of small between-trial variation ($\tau^2 = 0.1$). In the scenario when no interactions are present, both methods correctly detect no subgroups with enhanced treatment effect more than 91% of the time. When a single 1-way interaction is present, ie, there is a subgroup with enhanced treatment effect, both methods detect the correct subgroup around 77% of the time when a medium sized interaction is present in a sample size of 1000. The methods detect the majority of medium sized 1-way interactions when the sample size is ≥ 2500, and they detect the majority of large 1-way interactions for all sample sizes ≥ 1000. When two 1-way interactions are present, ie, 2 subgroups with enhanced treatment effect, the methods detect the correct subgroup the majority of the time when both interactions are either medium, large, or a combination of the 2 for all sample sizes ≥ 1000. Furthermore, the methods perform fairly well in detecting the correct subgroups when the sample size is 5000 and when one of the 1-way interactions is small and the second 1-way interaction is either medium or large.

When there is large between-trial variation ($\tau^2 = 0.9$), both methods perform very well when there are no interactions present; however, the modified SIDES method performs slightly better in this scenario. In general, the IPD-SIDES method clearly outperforms the modified SIDES method in the presence of large between-trial variation. Just to give an example, when 2 medium sized 1-way interactions are present with a sample size of 1000, the modified method detects the correct subgroups 51.2% of the time whereas the IPD-SIDES method detects the correct subgroups 84.8% of the time. The IPD-SIDES method performs very well when there are two 1-way interactions that are either medium or large in size for all sample sizes ≥ 1000. Moreover, the IPD-SIDES approach also performs well when one of the two 1-way interactions is small and the second 1-way interactions is either medium or large for a sample size of 5000.

TABLE 2 – Simulation results for modified SIDES and IPD-SIDES when there is small ($\tau^2 = 0.1$) and large ($\tau^2 = 0.9$) between-study variation and $N = 1000$. Results display % of correctly identified final trees

$T \times X_1$ Standardized Interaction Effect Size		$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.1$)				$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.9$)			
		None = 0	Small = 0.2	Medium = 0.5	Large = 0.8	None = 0	Small = 0.2	Medium = 0.5	Large = 0.8
None = 0	Modified SIDES	91.7	16.0	78.6	88.00	97.0	6.7	66.7	96.9
	IPD-SIDES	91.0	15.5	76.5	92.1	90.3	17.9	78.9	93.1
Small = 0.2	Modified SIDES	16.3	5.7	26.8	18.9	8.3	1.8	9.8	8.7
	IPD-SIDES	19.1	7.0	32.8	22.3	14.4	14.4	20.9	21.8
Medium = 0.5	Modified SIDES	77.6	23.9	74.3	82.2	70.1	9.0	51.2	67.0
	IPD-SIDES	65.9	23.4	74.6	89.8	77.1	23.0	84.8	91.6
Large = 0.8	Modified SIDES	94.1	30.4	90.1	100	97.5	16.1	76.9	97.1
	IPD-SIDES	92.4	22.5	92.8	100	92.1	22.1	81.6	100

TABLE 3 – Simulation results for modified SIDES and IPD-SIDES when there is small ($\tau^2 = 0.1$) and large ($\tau^2 = 0.9$) between-study variation and $N = 2500$. Results display % of correctly identified final trees

$T \times X_1$ Standardized Interaction Effect Size		$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.1$)				$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.9$)			
		None = 0	Small = 0.2	Medium = 0.5	Large = 0.8	None = 0	Small = 0.2	Medium = 0.5	Large = 0.8
None = 0	Modified SIDES	92.5	39.4	90.3	94.0	97.3	23.7	95.4	97.4
	IPD-SIDES	91.2	38.5	91.7	93.5	90.9	37.7	91.0	92.5
Small = 0.2	Modified SIDES	36.4	21.7	42.5	42.8	23.0	10.0	25.3	23.6
	IPD-SIDES	38.3	31.3	45.8	43.3	38.1	25.1	44.5	41.6
Medium = 0.5	Modified SIDES	92.6	47.8	99.6	99.1	93.8	23.1	97.3	99.3
	IPD-SIDES	91.7	43.3	99.9	99.7	93.0	44.1	92.5	99.5
Large = 0.8	Modified SIDES	93.9	41.4	99.9	100	98.1	24.0	98.5	100
	IPD-SIDES	93.7	43.4	99.8	100	92.7	44.5	99.9	100

TABLE 4 – Simulation results for modified SIDES and IPD-SIDES when there is small ($\tau^2 = 0.1$) and large ($\tau^2 = 0.9$) between-study variation and $N = 5000$. Results display % of correctly identified final trees

$T \times X_1$ Standardized Interaction Effect Size		$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.1$)				$T \times X_2$ Standardized Interaction Effect Size ($\tau^2 = 0.9$)			
		None = 0	Small = 0.2	Medium = 0.5	Large = 0.8	None = 0	Small = 0.2	Medium = 0.5	Large = 0.8
None = 0	Modified SIDES	93.0	60.8	94.3	88.5	97.6	56.4	98.3	97.1
	IPD-SIDES	91.4	62.0	93.2	83.6	90.5	70.2	93.0	83.9
Small = 0.2	Modified SIDES	69.7	71.3	74.3	76.0	53.9	40.4	64.1	56.3
	IPD-SIDES	69.9	63.5	75.1	79.1	61.1	63.4	78.6	81.7
Medium = 0.5	Modified SIDES	93.8	74.5	100	100	98.1	54.4	100	100
	IPD-SIDES	91.5	76.0	100	100	93.5	77.3	99.7	100
Large = 0.8	Modified SIDES	88.9	76.9	100	100	97.4	60.2	100	100
	IPD-SIDES	85.5	81.9	100	100	83.9	80.2	100	100

6 | APPLICATION TO BACK PAIN DATASET

The IPD-SIDES method was applied to the pooled acupuncture dataset described earlier with the aim of identifying subgroups with enhanced treatment effect that most benefit from acupuncture treatment for NSLBP. The splitting criterion was estimated using the fixed-effects model and then adjusted using the Sidak-based multiplicity adjustment to control for selection bias due to the inclusion of continuous covariates. The parameters had to be specified prior to applying the method. The minimum node size at any given time was set to 30. The maximum number of levels was set as being 4 ($L = 4$), the number of best splits considered for each node was set as 3 ($M = 3$), and the significance threshold was set to $V = 0.10$. The optimum relative improvement parameter values for each of the levels were determined using a grid search. The final subgroups were selected using the re-sampling procedure (1000 samples drawn).

6.1 | IPD-SIDES results

The results of the IPD-SIDES method are presented in Table 5. Three subgroups with enhanced treatment effect were identified for the change from baseline to short-term PCS outcome where the overall treatment effect was 3.75 (95% CI: 3.20, 4.30). Those with baseline MCS > 51.4 have a mean treatment benefit of 4.34 (95% CI: 3.43, 5.25), those with MCS > 51.4 and PCS ≤35.9 have an average treatment benefit of 5.44 (95% CI: 4.24, 6.63), and finally those participants with Age ≤ 43 have an average treatment benefit of 4.93 (95% CI: 3.90, 5.96). Hence, younger people or those with better mental

TABLE 5 – Subgroups identified by the IPD-SIDES method when applied to the pooled acupuncture dataset

Subgroups[a]	n	Treatment Effect (95% Confidence Interval, CI)	Interaction Effect	Unadjusted P-Value
Outcome: Short-term PCS				
Overall treatment effect (95% CI): 3.75 (3.20, 4.30)				
Candidate 1				
MCS > 51.4	1531	4.34 (3.43, 5.25)	0.94	0.086
MCS ≤ 51.4	2314	3.40 (2.72, 4.09)		
Candidate 2				
MCS > 51.4 and PCS ≤ 35.9	919	5.44 (4.24, 6.63)	2.38	0.016
MCS > 51.4 and PCS > 35.9	612	3.05 (1.90, 4.21)		
Candidate 3				
Age ≤ 43	1170	4.93 (3.90, 5.96)	1.69	0.005
Age > 43	2675	3.24 (2.59, 3.88)		
Outcome: Short-term MCS				
Overall treatment effect (95% CI): 2.50 (1.83, 3.17)				
Candidate 1				
MCS ≤ 54.5	2701	3.29 (2.47, 4.10)	2.65	0.001
MCS > 54.5	1144	0.64 (−0.21, 1.48)		

[a]The first row of each candidate subgroup is the selected subgroup with enhanced treatment effect. The second row is the disregarded subgroup.

functioning and worse physical functioning at baseline have better treatment benefit when the outcome is the change in physical functioning. One subgroup with enhanced treatment effect was identified for the change from baseline to short-term MCS outcome where the overall treatment effect was 2.50 (95% CI: 1.83, 3.17). Those with baseline MCS ≤ 54.5 have a greater average treatment benefit of 3.29 (95% CI: 2.47, 4.10). In other words, those with poorer mental functioning at baseline have greater treatment benefit when the outcome measure is the change in mental functioning.

7 | DISCUSSION

This paper proposes the IPD-SIDES method as a modified exploratory statistical approach for identifying subgroups in an IPD meta-analyses framework. The proposed method differs from the typical statistical interaction test approach to subgroup analyses in both a single trial and an IPD meta-analyses setting, thus overcoming some of the key concerns associated with the conventional approach. Although the development of this method was motivated by a research priority in the area of NSLBP, its application is not limited to this field. The proposed method can be applied as an exploratory tool in any other research discipline where subgroup identification in an IPD meta-analysis setting is of interest. The IPD-SIDES method is not an improvement or replacement for the standard non-exploratory approach, but is considered an additional exploratory approach for identifying potential subgroups to be used alongside standard methods. We recommend that any potential subgroups identified by the method be assessed to see if they are clinically or biologically plausible before testing them using standard methodology in further studies.

An IPD framework provides much improved statistical power and is thus ideal for subgroup analyses. The simulation studies demonstrated that the proposed method performs well in detecting subgroups in a number of scenarios, especially when there is large between-trial variation. The proposed IPD-SIDES method was also compared elsewhere using simulation studies to the extension of another relevant tree-based method called the Interaction Tree (IT)[15] where it was observed that the IPD-SIDES method was the more powerful method out the two.[37] A limitation of the proposed method however is that there is a possibility for the interaction tests being ecologically biased, which means that the observed across-study relationship estimated using a 1-stage approach may not be a true reflection of the individual level within-study relationships.[40] We did not explore this possibility when developing this method; however, it is something we aim to assess as further work.

For many years, a well-known limitation of tree-based methods is the issue of variable selection bias, meaning that the algorithm has a greater probability of selecting a covariate with a larger number of levels.[33,34] It is thus important for any tree method to ensure this bias is minimised. A number of approaches have been proposed that reduce this form of bias.[14,35] For example the GUIDE method uses a 2-stage approach for variable selection.[14] The first stage uses a chi-squared test for association between each covariate and outcome which ensures each covariate has the same chance of being selected under the null (ie, if having no predictive value). The variable with the most association (ie, smallest P-value) is selected. The second stage then searches the selected variable for the optimal split. The proposed IPD-SIDES method also reduces selection bias by utilising a Sidak-based multiplicity adjustment. This basically adjusts the splitting criteria P-value for covariates with many levels based on the number of levels it has, thus reducing selection bias.[16] For any continuous covariate, the method requires the selection of a cut-point as seen in the applied example. The objective of the method is to identify subgroups rather than to find optimal cut-points. The selection of the cut-point simply aids the subgroup search process. It may well be that the selection of the cut-point may be an artefact of the dataset, ie, it may not be reproducible and could possibly be spurious. However, the final step of the method implements a resampling-based procedure to assess reproducibility and remove any identified candidate subgroups that may be a spurious finding. Alternatively, cut-off values can be proposed by the clinician's priori to applying the method. Just to reiterate that if the method identifies potential subgroups, the next step would be to assess how clinically meaningful the subgroups and cut-points are before testing the subgroups in further studies.

The results of the simulation studies when using fixed-effects and mixed-effects models to estimate the splitting criterion were quite similar. From a computational standpoint, application of the method when using a fixed-effects model was much quicker to run than when using a mixed-effects model. Therefore, considering the extensive searching of splits required by the method, it would be computationally more efficient when working with extremely large datasets to estimate the splitting criterion using a fixed-effects model; otherwise, the mixed-effects model is preferred.

Like any meta-analysis study, it is important to make sure that the trials included are of a high quality to ensure the results and conclusions drawn are credible. Therefore, it is recommended that a risk of bias assessment is conducted for all potential trials prior to inclusion in an IPD meta-analyses. Furthermore, subgroup analysis in an IPD meta-analysis setting is only worthwhile if there is a commonality of covariates across studies. Thus, the trials considered for inclusion must also be checked at the covariate level to ensure that there are a reasonable number of covariates common to all the trials that can be explored for subgroups.

It has been demonstrated that the tree methods perform well when the outcome data are not normal. To give an example, Su et al were able to show that the IT method performed well in detecting interactions in a single trial setting when there were deviations from normality.[15] The simulation studies reported in this paper generated data assuming the outcome to be normally distributed and assuming the proportion of observations in each quadrant of the treatment-covariate interaction to be balanced. Therefore, future work should investigate how the IPD-SIDES method performs when the data deviate from normality, when there are varying degrees of imbalance in the data when forming a split, when different types of covariates are investigated, eg, ordinal or continuous covariates, and also when there are a varying total number of covariates. Furthermore, a significance threshold V was proposed and implemented into the IPD-SIDES algorithm to control for spurious subgroup detection when large or very large interactions are present. In this work, we only considered a threshold of $V = 0.10$ which seemed to work well; however, further work should investigate how much we can relax the significance threshold without affecting the performance of the method.

ORCID

Dipesh Mistry http://orcid.org/0000-0002-0875-9260

REFERENCES

1. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health Technol Assess (Winch Eng)*. 2001;5(33):1-56.

2. Lagakos SW. The challenge of subgroup analyses — reporting without distorting. *N Engl J Med*. 2006;354(16):1667-1669.

3. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*. 2005;365(9454):176-186.

4. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine — reporting of subgroup analyses in clinical trials. *N Engl J Med*. 2007;357(21):2189-2194.

5. Wang R, Ware JH. Detecting moderator effects using subgroup analyses. *Prevention Sci:Tthe official journal of the Society for Prevention Research*. 2011;

6. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline.

7. European Medicines Agency. Draft guideline on the investigation of subgroups in confirmatory clinical trials. *CHMP/539146/2013*. 2014.

8. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(1):99-119.

9. Tanniou J, van der Tweel I, Teerenstra S, KCB R. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC Med Res Methodol*. 2016;16(1):20

10. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall; 1984.

11. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30(24):2867-2880.

12. Dusseldorp E, Conversano C, Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. *J Computation Graphic Stat*. 2010;19(3):514-530.

13. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat Med*. 2014;33(2):219-237.

14. Loh W-Y, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med*. 2015;34(11):1818-1833.

15. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *SSRN eLibrary*. 2009.

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

17. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*. 2017;36(1):136-196.

18. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340:

19. Abdolell M, LeBlanc M, Stephens D, Harrison RV. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Stat Med*. 2002;21(22):3395-3409.

20. Keon Lee S. On generalized multivariate decision tree by using GEE. *Computation Stat Data Analysis*. 2005;49(4):1105-1119.

21. Lee S. On Classification and regression trees for multiple responses. In: Banks D, McMorris F, Arabie P, Gaul W, eds. *Classification, Clustering, and Data Mining Applications*. Heidelberg: Springer Berlin; 2004:177-184.

22. Loh W-Y, Zheng W. Regression trees for longitudinal and multiresponse data. 2013:495–522.

23. Sela R, Simonoff J. RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*. 2012;86(2):169-207.

24. Su X, Meneses K, McNees P, Johnson WO. Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *J R Stat Soc Ser C Appl Stat*. 2011;60(3):457-474.

25. Hajjem A, Bellavance F, Larocque D. Mixed effects regression trees for clustered data. *Stat Probabil Lett*. 2011;81(4):451-459.

26. Savigny P, Watson P, Underwood M. Early management of persistent non-specific low back pain: summary of NICE guidance. *BMJ*. 2009;338:b1805

27. Borkan JMMDP, Koes BP, Reis SMD, Cherkin DCP. A report from the second international forum for primary care research on low back pain: reexamining priorities. *Spine*. 1998;23(18):1992-1996.

28. Airaksinen O, Brox J, Cedraschi C, et al. Chapter 4 European guidelines for the management of chronic nonspecific low back pain. *Eur Spine J*. 2006;15(0):s192-s300.

29. Chou R, Qaseem A, Snow V, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. *Ann Intern Med* 2007;147(7):478–491.

30. Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. *Spine (Phila Pa 1976)*. 2014;39(7):618-629.

31. Ware JE Jr. SF-36 health survey update. *Spine (Phila Pa 1976)*. 2000;25(24):3130-3139.

32. Zhang H, Singer BH. *Recursive Partitioning and Applications.* 2nd ed. Springer; 2010.

33. Doyle P. The use of automatic interaction detector and similar search procedures. *Operation Res Quart (1970–-1977).* 1973;24(3):465-467.

34. Shih Y-S, Tsai H-W. Variable selection bias in regression trees with constant fits. *Computation Stat Data Analysis.* 2004;45(3):595-607.

35. Lausen B, Sauerbrei W, Schumacher M. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In: Dirschedl P, Ostermann R, eds. *Computational Statistics.* Heidelberg, Germany: Physika-Verlag; 1994:483-496.

36. Mistry D. Recursive partitioning based approaches for low back pain subgroup identification in individual patient data meta-analyses. PhD Thesis: Warwick Medical School, University of Warwick; 2014.

37. Mistry D. *Recursive Partitioning Based Approaches for Low Back Pain Subgroup Identification in Individual Patient Data Meta-Analyses.* Warwick Medical School, University of Warwick; 2014.

38. Abo-Zaid G, Guo B, Deeks JJ, et al. Individual participant data meta-analyses should not ignore clustering. *J Clin Epidemiol.* 2013;66(8):865–873.e864.

39. Hempel S, Miles JN, Booth MJ, Wang Z, Morton SC, Shekelle PG. Risk of bias: a simulation study of power to detect study-level moderator effects in meta-analysis. *Systemat Rev.* 2013;2:107

40. Hua H, Burke DL, Crowther MJ, Ensor J, Tudur Smith C, Riley RD. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Stat Med.* 2017;36(5):772-789.