Received: 27 November 2018 | Revised: 11 June 2019 | Accepted: 11 June 2019

DOI: 10.1002/sim.8308

RESEARCH ARTICLE

WILEY Statistics in Medicine

PSICA: Decision trees for probabilistic subgroup identification with categorical treatments

Oleg Sysoev[1] | **Krzysztof Bartoszek**[1] | **Eva-Charlotte Ekström**[2] | **Katarina Ekholm Selling**[2]

[1]Department of Computer and Information Science, Linköping University, Linköping, Sweden

[2]Department of Women's and Children's Health, Uppsala University, Akademiska Sjukhuset, Uppsala, Sweden

Correspondence
Oleg Sysoev, Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden.
Email: oleg.sysoev@liu.se

Present Address
Oleg Sysoev, Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden

Funding information
Swedish Research Council, Grant/Award Number: 2014-2161

Personalized medicine aims at identifying best treatments for a patient with given characteristics. It has been shown in the literature that these methods can lead to great improvements in medicine compared to traditional methods prescribing the same treatment to all patients. Subgroup identification is a branch of personalized medicine, which aims at finding subgroups of the patients with similar characteristics for which some of the investigated treatments have a better effect than the other treatments. A number of approaches based on decision trees have been proposed to identify such subgroups, but most of them focus on two-arm trials (control/treatment) while a few methods consider quantitative treatments (defined by the dose). However, no subgroup identification method exists that can predict the best treatments in a scenario with a categorical set of treatments. We propose a novel method for subgroup identification in categorical treatment scenarios. This method outputs a decision tree showing the probabilities of a given treatment being the best for a given group of patients as well as labels showing the possible best treatments. The method is implemented in an R package **psica** available on CRAN. In addition to a simulation study, we present an analysis of a community-based nutrition intervention trial that justifies the validity of our method.

KEYWORDS

bootstrap, decision trees, personalized medicine, random forest, subgroup discovery

1 | INTRODUCTION

It is very common that randomized trials are performed to investigate the efficiency of a new treatment. In these trials, a new treatment is compared to a control treatment, and if the new treatment is shown to be more efficient than the control it is suggested to be used on a population-wide level. Alternatively, in confirmatory subgroup analysis, effect of the treatment is investigated in the prespecified subgroups.[1]

Methods from personalized medicine[2] have drawn a lot of attention in medical and statistical literature.[3] These methods aim to identify and propose the best treatments to a patient with given characteristics (medical history). This clearly might lead to more efficient therapies than those proposed by confirmatory randomized trials. A lot of methods from personalized medicine are related to applications in genetics, ie, these methods detect treatments that persons with specific genetic biomarkers benefit of. From a statistical perspective, this typically reduces to a high-dimensional regression problem with binary input variables indicating the absence or presence of corresponding genetic biomarkers.

Another important category of personalized medicine is subgroup identification, a comprehensive survey of methods from this category is available in the literature.[4] The methods from this category identify subgroups of patients, which benefit from the same treatments, and this identification can be based on the characteristics of various natures (binary, categorical, real valued). Subgroup discovery methods can be applied to various experimental designs, including randomized clinical trials.[5]

Some personalized medicine methods are devoted to modeling optimal treatment regimes (OTR).[6-12] The primary purpose of these methods is to determine the optimal treatment for a given patient rather than detecting subgroups having similar treatment effects. While some methods for the OTR prediction are *black-box* models, many approaches were proposed to deliver interpretable optimal treatment decisions.[8-12] Compared to the subgroup identification methods, the OTR methods search for a single optimal treatment for a given patient or groups of patients.

We focus on the subgroup identification methods that are inspired by decision tree structures. Decision trees are easily interpretable, which makes them very convenient for policy making. A decision-maker is thus not only able to see what treatments are recommended but also which patient characteristics this recommendation is based on. Some comparative analysis of such methods is reported.[13]

It appears that the majority of subgroup identification methods focus on two-arm trials, ie, when the treatment set is binary (control/treatment). Methods such as QUINT,[14] Virtual Twins (VT),[5] Interaction Trees (IT),[15] and SIDES[16] are able to identify subgroups in the binary scenario. Being very efficient in some settings, all of these methods have peculiarities that in some situations can be considered as limitations. Most importantly, all these methods except QUINT are focused on finding the groups when the treatment is better than the control, but they ignore the situations when the inverse is true (called qualitative interaction). Among other peculiarities/limitations, one may mention inability of processing continuous outcomes (eg, VT), nonprobabilistic nature of the algorithm (eg, QUINT), overlapping subgroups (SIDES), providing information about the mean difference in outcome within a subgroup rather than stating the probability that one treatment is better than the other one (IT). A few methods go behind the binary scenario: recently, a method treating continuous treatments (ordered by dose) was proposed.[17]

When trials are performed with a categorical set of treatments, no subgroup identification method exists that aims at finding subgroups and predicting which set of treatments is the best. In principle, model-based (MOB) trees[18] can be used to explain the dependence of the outcome on the medical history variables (characteristics) and the treatment variable. However, because this method tries to explain the outcome itself rather than the dominance of some treatments (prognostic variable problem[19]), very long trees might be needed to identify necessary subgroups. This makes conventional MOB trees very hard to use in practical policy making. It is also possible to apply the Gi method[19] to a scenario with a categorical set of treatments, but this method outputs mean outcomes per treatment and subgroup rather than specifying the best treatments. It means that when two or more treatments have the same expected outcome, this method would not be able to identify such a situation due to randomness in the observed outcome mean.

We propose a novel method that is able to handle a scenario with categorical treatments (ie, when two or more different types of treatments are considered). We call this method Probabilistic Subgroup Identification for CAtegorical (PSICA) treatments. Our method is designed for randomized controlled trials and continuous outcome variables. We believe that it is of great importance for a subgroup identification method to provide statistical guarantees in the form of the probabilities of a treatment being the best for a given subgroup and, when data are not sufficient for a reliable decision, to state that there is no statistical guarantee that one of the treatments is more appropriate than the others. This differentiates our method from the OTR approaches and from many existing subgroup identification methods. Accordingly, our method first uses random forests to compute the probabilities that a treatment is the best for a given patient, and then these probabilities are summarized by a decision tree in which each terminal node shows probabilities for a treatment to be the best and the label showing most likely treatments. When all probabilities are large enough within a node, its label may contain all treatments, which is equivalent to saying “I don't know which treatment is the best” (ie, collect more data).

As an example, consider three treatments in which the outcomes are linear with respect to characteristics x. Figure 1 demonstrates such an example and some amount of observations corresponding to this setting. If the highest outcome implies the best treatment, treatment B is supposed to be the best for smaller values of x, treatment A should be the best for moderate x, and treatment C should be the best for the larger x. However, for smaller x, treatments A and B do not differ so much, which means that a subgroup discovery method would probably have a hard time to identify one best treatment. Figure 2 demonstrates the result of application of the PSICA method to these data. It clearly illustrates that the subgroups are identified in the way that was expected.

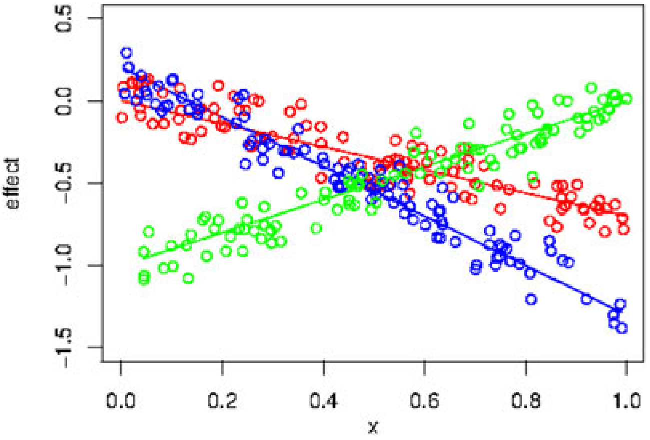

FIGURE 1 The outcomes for three treatments $\tau_1 = A$, $\tau_2 = B$, and $\tau_3 = C$ are generated as $y(x, \tau_1) = -0.7x + \epsilon$ (red), $y(x, \tau_2) = -1.5x + 0.2 + \epsilon$ (blue), $y(x, \tau_3) = x - 1 + \epsilon$ (green). Error term ϵ was generated as $N(0, 0.01)$ [Colour figure can be viewed at wileyonlinelibrary.com]

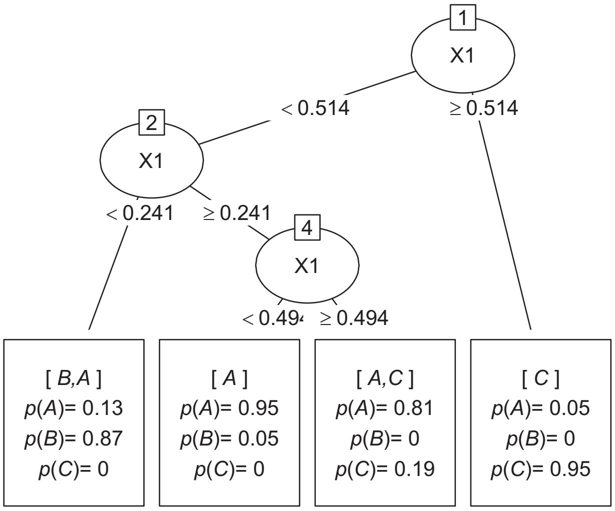

FIGURE 2 A PSICA tree showing subgroups, the probabilities of treatments being the best and labels containing the most likely optimal treatments

In our numerical experiments, we compare PSICA with QUINT when there are two treatments. We choose the QUINT method for comparisons because it is the only existing method not only capable of choosing the best treatment among two alternatives but also stating when the treatments are equivalent. In addition, we use PSICA to perform subgroup identification for the MINIMat trial[20] that was conducted in Matlab subdistrict, rural Bangladesh and contained 6 categorical interventions (treatments).

In Section 2, we present the PSICA method. In Section 3, we present our numerical simulations and a real case study. Section 4 contains conclusions and discussion.

2 | PSICA TREES

The problems of subgroup identification and some notation are introduced first. Given a data set $D = \{(X_i, Y_i, t_i), i \in 1, \dots, n\}$, where $X_i = (X_{i1}, \dots, X_{ip})$ is a set of characteristics (input variables, predictors) for patient i, Y_i is the outcome of the given treatment t_i, where t_i is one of the treatments that belong to the set $\mathcal{T} = \{\tau_1, \dots, \tau_m\}$. We assume that X_i values were obtained as a realization of a random variable x with components $x^1, \dots, x^p$. The response $Y(x, \tau)$, called a *potential outcome* (or simply outcome), is an outcome of a given treatment τ, and we assume that all treatments are possible to use for any patient. In practice, a patient with some characteristics X_i is assigned to only one of the treatments t_i, and the outcome Y_i is observed. The remaining $Y(x, \tau)$, $x = X_i$, $\tau \neq t_i$ are normally not available. However, the observed outcome (X_i, Y_i, t_i) is related to the potential outcomes as

$$Y_i = \sum_{j=1}^{m} Y(X_i, \tau_j) \cdot I(\tau_j = t_i),$$

where $I(z)$ is equal to one when z is true and zero otherwise.

In randomized controlled trials, the probabilities of assigning a patient with characteristics x to different treatments do not depend on these characteristics. Assuming in addition that the treatment status of a patient does not affect potential outcomes of other patients and that there are no hidden versions of the treatments,[21] the expected potential outcome becomes equal to the expected observed outcome per treatment, ie, $E(Y_i|x = X_i, \tau = t_i) = E(Y(X_i, t_i))$.

We assume that $Y(x, \tau) = f(x, \tau)+\epsilon$, where $f(x, \tau)$ is the expected potential outcome for a given x and τ. In agreement with the previous assumptions, the error terms ϵ are assumed to be independent between the patients and also independent between different treatment options of the same patient. The input variables may be categorical, ordinal or real valued, and the outcome is considered to be real valued.

In a binary setting, ie, when $\mathcal{T} = (\tau_1, \tau_2)$, the subgroup identification problem can be defined as finding subgroups G such that

$$\pi(G, \tau_2, \tau_1) = p\left(Y(x, \tau_2) > Y(x, \tau_1) \mid x \in G\right) > 1 - \alpha,$$

where α is some risk level, eg, 0.05. This means that it is of interest to find subgroups of patients for which the second treatment is significantly better than the first one (which typically is a control treatment). Another interesting scenario is a qualitative subgroup identification, which means that the interesting subgroups are either those having $\pi(G, \tau_2, \tau_1) > 1-\alpha$ or those satisfying $\pi(G, \tau_1, \tau_2) > 1 - \alpha$.

When there are more than two treatments, the subgroup identification problem can be defined as follows: identify groups G and subsets of treatments $T \subset \mathcal{T}$ such that

$$p\left(Y(x, \tau') > Y(x, \tau'') \mid x \in G, \tau' \in T, \tau'' \in \mathcal{T} \setminus T\right) > 1 - \alpha.$$

It means that we want to either identify which treatments are useful and can be prescribed to a patient (treatments from T) or which treatments are useless for this subgroup and should not be given to these patients (treatments from $\mathcal{T}\setminus T$). Note that we require $T \neq \mathcal{T}$ because otherwise $\mathcal{T} \setminus T$ becomes empty and (X, τ'') becomes undefined.

The PSICA trees partition the input space into nonoverlapping regions and provide a label for each region and a probability distribution on the set $\{\tau_1, \dots, \tau_m\}$ specifying how likely it is that a given treatment is the best one for the group of patients characterized by the input values from this region. The PSICA tree computation consist of two steps: estimation of distributions and growing the PSICA tree.

The first step of PSICA tree computation implies estimating distributions $\pi_k(x)$, which is a probability that the treatment τ_k is better than all alternative treatments for a given x. To estimate $\pi_k(x), k = 1, \dots, m$ by simulation, we need to be able to generate samples from the joint distribution $p(Y(x, \tau_1), \dots, Y(x, \tau_m))$. This distribution shows how likely it is that if a patient with characteristics x receives treatment τ_1, then the outcome will be $Y(x, \tau_1)$, and if the same patient receives τ_2, then the outcome will be $Y(x, \tau_2)$, etc. If we are able to generate some number of samples $Y^b = (Y_1^b(x), \dots, Y_m^b(x)), b = 1, \dots, B$ from this distribution, then $\pi_k(x)$ can be estimated as

$$\pi_k(x) = \frac{1}{B}\sum_{b=1}^{B} I\left(Y_k^b(x) > \max_{j=1,\dots,m,j\neq k} Y_j^b(x)\right). \tag{1}$$

To generate samples from $p(Y(x, \tau_1), \dots, Y(x, \tau_m))$, we divide data D into subsets $D_k = \{(X_i, Y_i, t_i) : t_i = \tau_k, (X_i, Y_i, t_i) \in D\}$ for all $k = 1, \dots, m$. Each subset D_k corresponds to one of the treatments. The partitioned data are further used to generate samples Y^b by method 1 or method 2.

Method 1 implies that B data sets $D_k^b, b = 1, \dots, B$ are constructed by bootstrapping observations from D_k, we denote it as $D_k^b \sim \text{Bootstrap}(D_k)$. Afterwards, a machine learning model $M_k^b(x)$ is fit to each D_k^b by using y as response variable and x as the set of predictor variables. We propose to use conditional inference random forest[22] models but in principle any other machine learning (regression) model can be employed. Finally, samples $Y^b = (Y_1^b(x), \dots, Y_m^b(x))$ for any given x are generated as $Y_k^b(x) = M_k^b(x), k = 1, \dots m, b = 1, \dots, B$.

Method 2 implies fitting a machine learning model to each D_k, estimating the prediction $M_k(x)$ and then estimating the variance $\sigma_k^2(x)$ of prediction by using the bias-corrected infinitesimal jackknife,[23] see formula (7) in the corresponding paper. Finally, components $Y_k^b(x)$ of the samples are generated from a normal distribution with mean $M_k(x)$ and variance $\sigma_k^2(x)$ for each $b = 1, \dots, B$. Estimation of $\pi_k(x)$ is summarized in Algorithm 1.

Algorithm 1 Computation of the distributions of the treatment effects

```
Given $D = \{(X_i, Y_i, t_i), i \in 1, \dots, n\}, t_i \in \mathcal{T} = \{\tau_1, \dots, \tau_m\}$, method = 1 or 2 .
for $k = 1$ to $m$ do
    Compute $D_k = \{(X_i, Y_i, t_i) : t_i = \tau_k, (X_i, Y_i, t_i) \in D, i = 1, \dots, n\}$
end for
if method = 1 then
    for $b = 1$ to $B$, $k = 1$ to $m$ do
        $D_k^b \sim \text{Bootstrap}(D_k)$.
        Compute $M_k^b(x)$ from $D_k^b$
        Compute $Y_k^b(X_i) = M_k^b(X_i)$ for each $(X_i, Y_i, t_i) \in D$
    end for
end if
if method = 2 then
    for $k = 1$ to $m$ do
        Compute $M_k(x)$ and $\sigma_k^2(x)$ from $D_k$
        Generate $Y_k^b(X_i) \sim N(M_k(X_i), \sigma_k^2(X_i))$ for each $b = 1, \dots, B$,
            for each $(X_i, Y_i, t_i) \in D$
    end for
end if
for $i = 1$ to $n$, $k = 1$ to $m$ do
    Compute $\pi_k(X_i)$ by using (1)
end for
```

In order to make a choice between method 1 and method 2 in a practice, the following arguments can be considered. Method 1 is based on the bootstrap principle, which aims to approximate the sampling process from the true data generating distribution by sampling from the data set instead. This might lead to biased estimates and high costs in terms of computational time. However, when the true outcome distribution is not Gaussian, method 1 may still be much less biased than method 2, which assumes the normality of the outcome distribution.

The second step of PSICA tree computations implies growing a tree summarizing the probabilities $\pi_k(X_i)$ in such a way that interesting subgroups are discovered. We suggest two alternative methods for the tree growing process: method A and method B. Method A requires growing a large tree and then letting a user to prune it until interpretable applied results are achieved and at the same time, the tree becomes small enough to be used for policy making. Random forests are known to be very flexible models that are robust to overfitting,[24] which means that increasing the number of trees in the forests usually decreases the bias in estimation of the outcomes without increasing the variance. At the same time, increasing the number of bootstrap samples B decreases the variance in $\pi_k(x)$. Accordingly, for sufficiently large number of trees in the forests and for a sufficiently high number of bootstrap samples B, the probabilities $\pi_k(x)$ are expected to have small bias and small variance. However, it can be hard to judge whether the settings used by a user are sufficiently large, which means that there might be a risk for producing spurious subgroups. To remedy this problem, we suggest method B, which implies early pruning of the tree (prepruning) that guarantees that fewer spurious results are detected. However, since this method is based on hypothesis testing, there is a risk that some interesting subgroups are not found.

Method A employs standard decision tree growing principles.[25] More specifically, a data set $\Delta_0 = \{(X_i, P_i), P_i = (\pi_1(X_i), \dots, \pi_m(X_i))\}$ with inputs X_i and a vector response P_i is constructed first. This data set is partitioned recursively by using various binary splitting rules R_j (constructed differently for real-valued and categorical split variables) until some stopping criterion is met. This criterion might include constraints on the minimal amount of the observations in the node, maximal tree depth, and other criteria. To decide which splitting rules need to be used, the data set Δ that corresponds to a tree node before split R_j and the data sets after this split Δ_1 and Δ_2 are considered, and loss function values $L_1 = L(\Delta)$, $L_2 = L(\Delta_1)$, and $L_3 = L(\Delta_2)$ are computed. A splitting rule that maximizes *information gain*

$$g(\Delta, R_j) = L_1 - (L_2 + L_3) \tag{2}$$

is chosen to split the current node.

When no further split can be done, labels are assigned to the terminal nodes. In our settings, the following summary might be presented for a tree leaf corresponding to a data set Δ:

- Aggregated probabilities of each treatment being the best

$$\pi_k(\Delta) = \frac{1}{|\Delta|} \sum_{(X_i,P_i)\in\Delta} \pi_k(X_i), \tag{3}$$

  where $|\Delta|$ denotes the number of observations in Δ.
- A set of *useless treatments*$\mathcal{T}_u$. The probabilities $\pi_k(\Delta)$ are sorted in increasing order as $(\pi_{k_1}(\Delta), \dots, \pi_{k_m}(\Delta))$ and m' is found such that $\sum_{i=1}^{m'} \pi_{k_i}(\Delta) \le \alpha$ and $\sum_{i=1}^{m'+1} \pi_{k_i}(\Delta) > \alpha$, where α is a risk level (eg, $\alpha = 0.05$). The set $\mathcal{T}_u$ is computed as $\mathcal{T}_u = \{\tau_{k_1}, \dots, \tau_{k_{m'}}\}$.
- A set of *potential treatments*

$$\mathcal{T}_p = \mathcal{T} \setminus \mathcal{T}_u. \tag{4}$$

To enable successful subgroup identification, an appropriate loss function needs to be selected. To identify an appropriate function, it is important to consider how the resulting tree is going to be used in decision-making. We assume that after a decision-maker assigns the patient into one of the terminal nodes of the decision tree, the aggregated probabilities $\pi_k(\Delta), k = 1, \dots, m$ are compared, and the treatments from $\mathcal{T}_u$ will be excluded by the decision-maker. The remaining treatments are the potential treatments, and ideally, a further investigation of which one of them should be prescribed to a given patient will be performed. However, it is also very likely that the aggregated probabilities corresponding to treatments from $\mathcal{T}_p$ will be used by the decision-maker directly as an indicator of which treatment should be used.

Therefore, we define the loss function $L(\Delta)$ as the cost of assignment of the individuals represented by Δ to the treatments that they do not benefit from. More specifically, we define *truncated* probabilities as

$$\begin{aligned} \hat{\pi}_k(\Delta) &= \frac{\pi_k(\Delta)}{\sum_{i\in\mathcal{T}_p} \pi_i(\Delta)}, k \in \mathcal{T}_p \\ \hat{\pi}_k(\Delta) &= 0, k \notin \mathcal{T}_p, \end{aligned} \tag{5}$$

and therefore the cost of classifying a patient to a wrong treatment is

$$\sum_{k=1}^{m} \sum_{j\in\mathcal{T}_p, j\neq k} c_{kj} p(\text{Assigned to } \tau_j \text{ given } \tau_k \text{ is best}) \cdot p(\tau_k \text{ is best}), \tag{6}$$

where $\{c_{kj}, k, j = 1, \dots, m\}$ are costs of giving the patient treatment τ_j given that his/her best treatment is τ_k. A simple set of cost values is a zero-one cost: $c_{kj} = 1$ when $k \neq j$ and zero otherwise.

Equation (6) can be rewritten as

$$\sum_{k=1}^{m} \sum_{j\in\mathcal{T}_p, j\neq k} c_{kj} \hat{\pi}_j(\Delta) \cdot \pi_k(x)$$

and, by summing up the loss values for the observations within Δ, we obtain the following loss function:

$$L(\Delta) = \sum_{(X_i,P_i)\in\Delta} \sum_{k=1}^{m} \sum_{j\in\mathcal{T}_p, j\neq k} c_{kj} \hat{\pi}_j(\Delta) \cdot \pi_k(X_i). \tag{7}$$

If the zero-one loss is used, it is easy to show that (7) can be simplified as

$$L(\Delta) = |\Delta| \sum_{k=1}^{m} \pi_k(\Delta) \cdot (1 - \hat{\pi}_k(\Delta)). \tag{8}$$

Method B involves early stopping to avoid discovery of spurious subgroups. The tree growing procedure is identical to the first approach described above with the only exception that the information gain function g is modified in order to avoid splits that may generate spurious subgroups. More specifically, the modified information gain g' is defined as

$g'(\Delta, \Delta_1, \Delta_2) = g(\Delta, \Delta_1, \Delta_2) \cdot G(\Delta_1, \Delta_2)$, where $G(\Delta_1, \Delta_2)$ is equal to one if the distributions $\Pi_1 = \{\pi_k(\Delta_1), k = 1, \dots, m\}$ and $\Pi_2 = \{\pi_k(\Delta_2), k = 1, \dots, m\}$ differ significantly and zero otherwise.

To compute function G, we perform a chi-square test where we compare Π_1 and Π_2. For each Π_j, we compute counts

$$\left\{ n_{kj} = \left\lceil \pi_k(\Delta_j) \cdot |\Delta_j| \cdot \omega_j \right\rceil, k = 1, \dots, m \right\}, \tag{9}$$

where ω_j is an inflation factor defined as the standard deviation of the uniform distribution $U[0, 1]$ (which is equal to $1/\sqrt{12}$) divided by the standard deviation of $\{\pi_k(X_i) : (X_i, P_i) \in \Delta_j\}$. The purpose of the inflation factor is to give higher weights to the distributions of $\pi_k(X_i)$ that have low variance (and, thus, more confident). After the counts for Π_1 and Π_2 are computed, these counts are combined into a two-way table, and the standard chi-square test is performed. If its p-value p_χ is lower than a risk level α, we set $G = 1$ otherwise $G = 0$.

The summary of the PSICA tree growing strategy is given in Algorithm 2.

Algorithm 2 Computation of PSICA tree

```
Given $\Delta_0 = \{(X_i, P_i) : P_i = (\pi_1(X_i), \dots, \pi_m(X_i)), i = 1, \dots, n\}$, method = $A$ or $B$, risk level $\alpha$.
Output: SplitNode($\Delta_0$).
function GETLOSS($\Delta$)
    Compute $\pi_1(\Delta), \dots, \pi_m(\Delta)$ by using (3).
    Compute $\hat{\pi}_1(\Delta), \dots, \hat{\pi}_m(\Delta)$ by using (5)
    Compute $L(\Delta)$ by using (8)
    Output: $L(\Delta)$
end function
function GETMASK($\Delta_1, \Delta_2, \alpha$)
    Compute $\{\pi_i(\Delta_j), i = 1, \dots, m, j = 1, 2\}$ by applying (3).
    Compute $n_{kj}$ by using (9), $k = 1, \dots, m, j = 1, 2$
    Compute $p$-value $p_\chi$ based on $\{n_{kj}, k = 1, \dots, m, j = 1, 2\}$.
    Set $G = 1$ if $p_\chi \le \alpha$ and $G = 0$ otherwise
    Output: $G$
end function
function SPLITNODE($\Delta$)
    if Stopping criterion is met for $\Delta$ then
        Output: NULL
    else
        for $j = 1$ to $p$, each $R_j$ do
            Split $\Delta = \Delta_1 \cup \Delta_2$ by using $R_j$.
            for each $(\Delta_1, \Delta_2)$ do
                Compute $g(\Delta, R_j)$ = getLoss($\Delta$) − getLoss($\Delta_1$) − getLoss($\Delta_2$)
                if method = $B$ then
                    $g(\Delta, R_j) \leftarrow g(\Delta, R_j) \cdot$ getMask($\Delta_1, \Delta_2, \alpha$)
                end if
            end for
        end for
        Compute $R = \arg\max_{j,R_j} g(\Delta, R_j)$
        Output: $R$, $\Delta_1$, $\Delta_2$ , Splitnode($\Delta_1$) and Splitnode($\Delta_2$).
    end if
end function
```

3 NUMERICAL EXPERIMENTS

Our PSICA method was implemented in an R package ***psica***, which is available on CRAN.[26] To evaluate the efficiency of the method, we tested it with the following models:

$$\text{M1} : y(x, \tau) = (2\text{th}(2x) + 3)I(\tau = \tau_1) + +(2\text{th}(x) + 2.3)I(\tau = \tau_2) + \epsilon, \tag{10}$$

where $\epsilon \sim N(0, 0.8^2)$ and th(x) is the hyperbolic tangent function. The variance in this and the following models was adjusted in such a way that the highest signal-to-noise ratio is approximately 10. Some properties of this function considered on the interval $[-1, 1]$ are that the function is relatively complex (ie, includes nonlinearities) and that τ_1 is best in the entire interval while the effect of τ_1 and τ_2 becomes very similar around $x = -0.5$. Therefore, one can expect that subgroup identification methods should be able to either identify τ_1 as the best treatment or they should be uncertain, for example, around $x = -0.5$ and especially for smaller data sets. The QUINT method is aimed at finding qualitative interactions, ie, it assumes that there exist regions where τ_1 is better than τ_2 and other regions where τ_2 is better than τ_1. It means that this method is expected to fail in finding such interactions when the data are generated from M1.

$$\text{M2/M3}: \; y(x,\tau) = 0.5I(x_1 \geq 0 \text{ and } x_2 \geq 0)I(\tau = \tau_1) + 0.5I(x_1 < 0 \text{ and } x_2 < 0)I(\tau = \tau_2) + \epsilon, \tag{11}$$

where $\epsilon \sim N(0, 0.2^2)$ (M2) and $\epsilon \sim \text{Laplace}(0, 0.2^2)$ (M3). These models contain qualitative interactions that are expected to be discovered by QUINT and also can be used to compare the effect of the error distribution (normal vs Laplace) on the performance of subgroup identification methods.

$$\text{M4}: \; y(x,\tau) = \sum_{i=1}^{40} x_i + 5x_1 I(x_1 > 0.5)I(\tau = \tau_1) + 5I(x_1 < 0.5 \text{ and } x_2 > 0.5)I(\tau = \tau_2) + \epsilon, \tag{12}$$

where $\epsilon \sim N(0, 2^2)$. This model is interesting to consider because it involves many variables in creating the main effect and a few variables that interact with the treatments.

$$\text{M5}: \; y(x,\tau) = (-0.7x - 0.7)I(\tau = \tau_1) + (-1.5x - 1.1)I(\tau = \tau_2) + (x - 1)I(\tau = \tau_3) + \epsilon, \tag{13}$$

where $\epsilon \sim N(0, 0.2^2)$. Model M5 is similar to the model explained in Figure 1. It contains three treatments and it can thus not be processed by binary subgroup identification methods like QUINT. However, this model is good enough to study the behavior of PSICA model in a simple setting.

$$\text{M6}: \; y(x,\tau) = \sum_{i=1}^{40} x_i + 5x_1 I(x_1 > 0.5)I(\tau = \tau_1 \text{ or } \tau = \tau_2) + 10(x_1 < 0 \text{ and } x_0 =' K1')I(\tau = \tau_3) + \epsilon, \tag{14}$$

where $\epsilon \sim N(0, 2^2)$, and $\mathcal{T} = \{\tau_1, \ldots, \tau_4\}$. In this model, there is a main effect and also complex treatment effects: one subgroup benefits from treatments τ_1 and τ_2 while another subgroup benefits from treatment τ_3. None of the patients benefits from τ_4. This model also includes a categorical variable x_0 with four unique values, and this variable is important in defining one of the subgroups. This model can thus be regarded as good test of PSICA trees in real complex scenarios.

We perform the following numerical experiments 200 times. First, we generate data from models M1 to M6 with n observations, where $n = 300, 900$, or 1800 and a randomized treatment assignment, where each x component is generated as $U[-1, 1]$. To make the correct subgroup identification even more difficult for the estimation algorithms, we add a number of irrelevant input variables generated as $U[-1, 1]$ to each data set: two variables for models M1, M2, M3, and M5, 160 variables for M4 and M6. In the next step, we perform subgroup identification by using PSICA (for M1 to M6) and QUINT (for M1 to M4). When computing PSICA trees, we use three alternatives: method m_1 denotes PSICA trees with probabilities computed by the bias-corrected infinitesimal jackknife (method 2 in Algorithm 1) and the number of variables per split in the random forest equals the total amount of input variables p, method m_2 denotes PSICA trees with probabilities computed by the bias-corrected infinitesimal jackknife and the number of variables per split in the random forest equal to $\sqrt{p}$, method m_3 PSICA trees computed by the bootstrap approach (method 1 in Algorithm 1) and the number of variables per split in the random forest equal to $\sqrt{p}$. Method m_3 is computed only for $n = 300$ due to high computational burden. Other settings were specified as $B = 500$, all PSICA trees used method B (see Algorithm 2) with $\alpha = 0.05$, number of trees in a forest equal to 100, minimal amount of observations for splitting the node in a tree equal to $n/10$ in the trees belonging to forests and $n/5$ in the PSICA trees. When computing QUINT trees (denoted as method m_4), we use the bootstrap pruning and default settings specified in the corresponding R package.[27]

The performance of the methods is evaluated by computing the following metrics: accuracy (a), uncertainty (u), and suspect (s). Given that for each feasible x, a tree delivers the predicted best treatments $\mathcal{T}_p$ while the true best treatments are T_p, metrics a and u are defined as follows:

$$a(D) = \frac{1}{n} \sum_{(X,Y,T)\in D} I(T_p(X) \subseteq \mathcal{T}_p(X)) \tag{15}$$

$$u(D) = \frac{1}{n} \sum_{(X,Y,T)\in D} I\left(|\mathcal{T}_p(X)| > |T_p(X)|\right), \tag{16}$$

where $|S|$ denotes the number of elements in a set S. Accuracy values represent proportions of the correct predictions while the uncertainty values specify how uncertain the tree is. Note that a tree can in principle achieve 100% accuracy by predicting all treatments as a full set $\mathcal{T}$, but it will also imply 100% uncertainty.

The *suspect* $s(\Delta)$ is defined as a the sum of the amounts of observations corresponding to the nodes that are immediately above the irrelevant splits divided by the sum of the amounts of observations corresponding to all nodes in the tree. Therefore, if an irrelevant variable is located in the top levels of the tree, the suspect value is expected to be high.

For PSICA trees, we also compute a measure, which we call *decision accuracy*. As it was discussed in Section 2, we assume that when a terminal node in the PSICA tree returns a set of potential treatments $\mathcal{T}_p$, and this set contains more than one treatment, a decision-maker is ideally supposed to make further investigations regarding which of these treatments should be given to a patient. However, it is also likely that the decision-maker will use the aggregated probabilities shown in the corresponding tree node to make a decision. However, this might not be a good strategy in some situations. Suppose $\mathcal{T} = \{\tau_1, \tau_2\}$ and in the given tree node $\pi_1 = 0.45$ and $\pi_2 = 0.55$. Although treatment τ_2 has a somewhat higher probability, it is clear that the model is quite unsure about which treatment is the best one for the group of patients associated with the given tree node. This means that, in this case, a further investigation is probably the most reasonable option. Assume though that the PSICA tree returns a set of truncated probabilities $\{\hat{\pi}_k(x), k = 1, \dots, m\}$ for a given x and the decision-maker makes a decision as $\tilde{\tau}(x) \sim Multinomial\,(\hat{\pi}_1(x), \dots, \hat{\pi}_m(x))$. Decision accuracy measures the proportion of the correct decisions in this scenario as

$$\delta(D) = \frac{1}{n} \sum_{(X,Y,T)\in D} I\left(\tilde{\tau}(X) \in T_p(X)\right). \tag{17}$$

Figures 3, 4, 5, and 6 illustrate the results of our simulation experiments, and the underlying data tables are provided in the Appendix, see Tables A1, A2, A3, and A4.

It can be concluded that m_1, m_2, and m_3 provide a similar accuracy, which is close to 100% in the majority of scenarios, both when binary treatments are used and when categorical scenarios are considered. Method m_4 (QUINT) has lower accuracy values, especially for smaller data and when there are many irrelevant variables (model M4). Accuracies of models m_1, m_2, m_3 also decrease when the data model is complex and there are many irrelevant predictors (model M6). By comparing the accuracies of methods m_1-m_3 across M2 and M3, no noticeable difference can be detected, which indicates that PSICA trees are not so sensitive to the error distribution.

When comparing uncertainties, an interesting fact can be observed: allowing the conditional inference random forest to use all input variables at any split (method m_1) leads to lower uncertainty rates than the setting $\sqrt{p}$ variables at any split (method m_2), which is recommended in the literature for the random forests. This happens because the trees in the forests are grown by means of early stopping involving hypothesis testing: if there are no relevant variables in the randomly selected subset of input variables, the corresponding tree node will not be split further. It may lead to deficient trees in some cases. It can also be observed that uncertainty rates decrease with increasing sample size for model m_1, while for models m_2 and m_4, these rates usually do not change much or sometimes increase. Noticeably, uncertainty rates of m_1 are generally lower than those of m_2 and m_4, and the uncertainty rates of m_2 are generally comparable to the rates of m_4 with the two exceptions. The first exception is a simple model (M1) where m_4 has high uncertainty rates. The second exception is a complex model (M4) where the m_2 rates are lower than the m_4 rates for larger n. The uncertainty rates of m_3 are often lower than those of m_1 indicating that applying bootstrap instead of the variance approximation might lead to better decisions. The price of this is a much higher computational time. Figure 5 illustrates that PSICA is good in finding relevant predictors: the suspect rates change approximately between 0 and 0.1. QUINT gives relatively low suspect rates for models M2 and M3. However, when a complex data model with many irrelevant predictors is processed (model M4), m_4 appears to have problems in finding the relevant predictors. When the data model implies that one treatment is the best one for all observations (model M1), QUINT produces sometimes trees with a root node only (which are excluded from computations of the suspect) and sometimes produces trees with irrelevant splits. When $n = 1800$, all computed QUINT trees contained only the root node, and therefore no suspect value is reported.

Figure 6 demonstrates that the decision accuracies for method m_1 are often very high (0.8-1.0), and they grow with increasing sample size. Method m_2 has somewhat lower decision accuracies, which confirms our previous finding: using all input variables at any split leads to a better performance of PSICA trees. Decision accuracies generated by method m_3 are comparable to the results obtained by m_1.

To study the effect of the risk level α in the tree pruning algorithm (method B), we perform additional 400 simulation experiments. In each experiment, we randomly choose n from the set $\{300, 900, 1800\}$ and the data model from $\{M1, \dots, M6\}$. A data set of size n is then generated according to the selected data model, and the generated data set is processed by the method m_1, and accuracy, uncertainty, suspect, and decision accuracy values are computed. Method m_1 was chosen because it had an overall high accuracy, relatively low uncertainty and low suspect rates in the previous experiments. The results are illustrated by Figure 7, and more detailed information is provided in Table A5.

The accuracy and decision accuracy rates were roughly the same for different α values, but the uncertainty and the suspect rates were more affected by the choice of α. Figure 7 illustrates that the suspect rates tend to increase when α increases. At the same time, uncertainty rates tend to decrease when α increases, which means that too low values of α, eg, $\alpha = 10^{-5}$, are not optimal either. The value $\alpha = 0.05$ has relatively low uncertainty rates and relatively low suspect rates and thus can be recommended in practice.

In addition to the simulation experiments, we analyze the so called MINIMat data[20] with the PSICA method. The MINIMat trial was conducted in the Matlab subdistrict, rural Bangladesh. In this area, 4436 pregnant women were enrolled between November 2001 and October 2003 to take part in the trial. The design and interventions of the MINIMat trial have previously been described in detail.[28] Very briefly, pregnant women were individually and randomly allocated in a 2 by 3 factorial design into two prenatal food and three micronutrient supplementation groups. Food supplementation was promoted to start either in early pregnancy (E for early) or at the women's own liking (U for usual). The three micronutrient groups were: 30 mg of iron supplementation (X), 60 mg iron (Y), 30 mg of iron, 400 mg of folic acid, and 13 other micronutrients (Z). At enrollment and during pregnancy, characteristics of the women and their households were collected. In this example, 124 variables, such as maternal anthropometry, parity, education, morbidity, exposure to domestic violence, as well as household food insecurity and assets, during the time of pregnancy were included as inputs.

The outcome variable is the children's height-for-age z-score at 54 months (HAZ), and the aim of PSICA tree analysis is to find out which interventions increase HAZ of the children. We computed a PSICA tree with pruning, $B = 1000$,

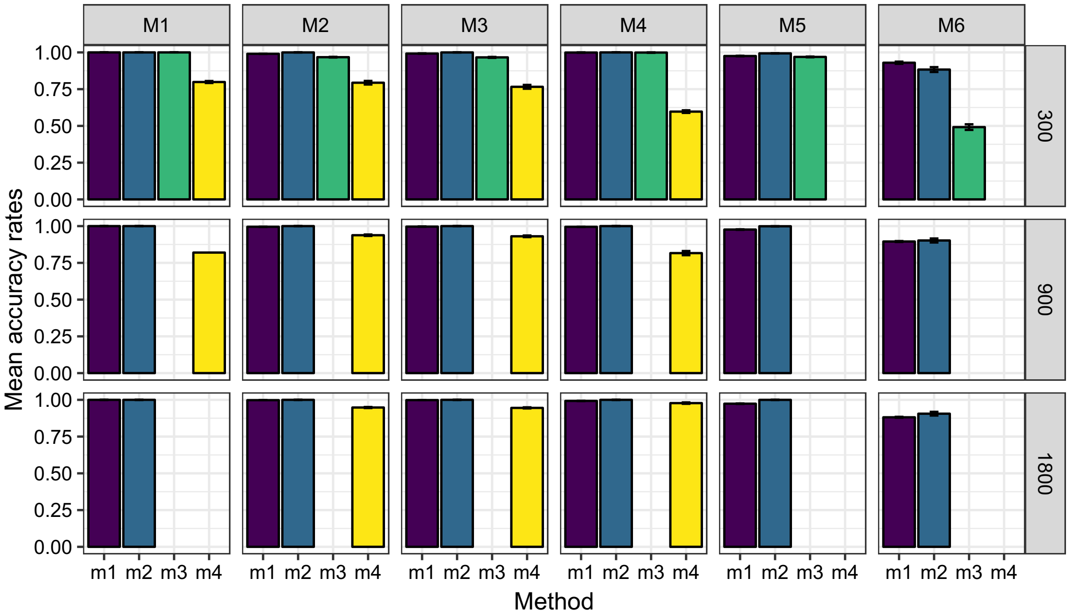

FIGURE 3 Mean accuracy rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is specified by the whiskers. Each panel of the graph corresponds to some data model (defined by the column title) and some data size (defined by the row title) [Colour figure can be viewed at wileyonlinelibrary.com]

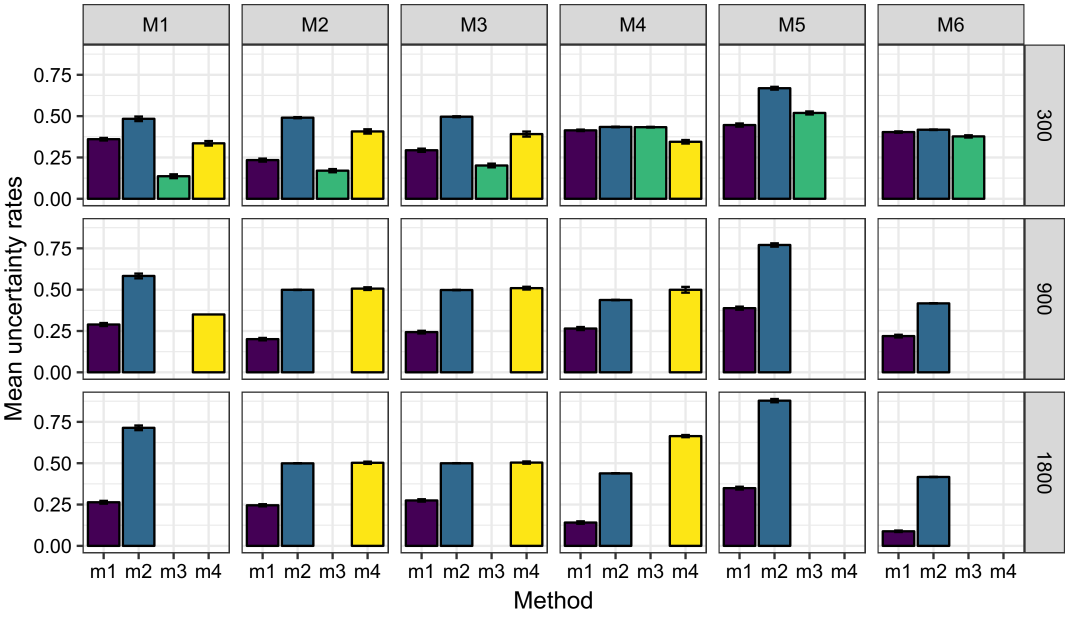

FIGURE 4 Mean uncertainty rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is represented by the whiskers. Each panel of the graph corresponds to some data model (defined by the column title) and some data size (defined by the row title) [Colour figure can be viewed at wileyonlinelibrary.com]

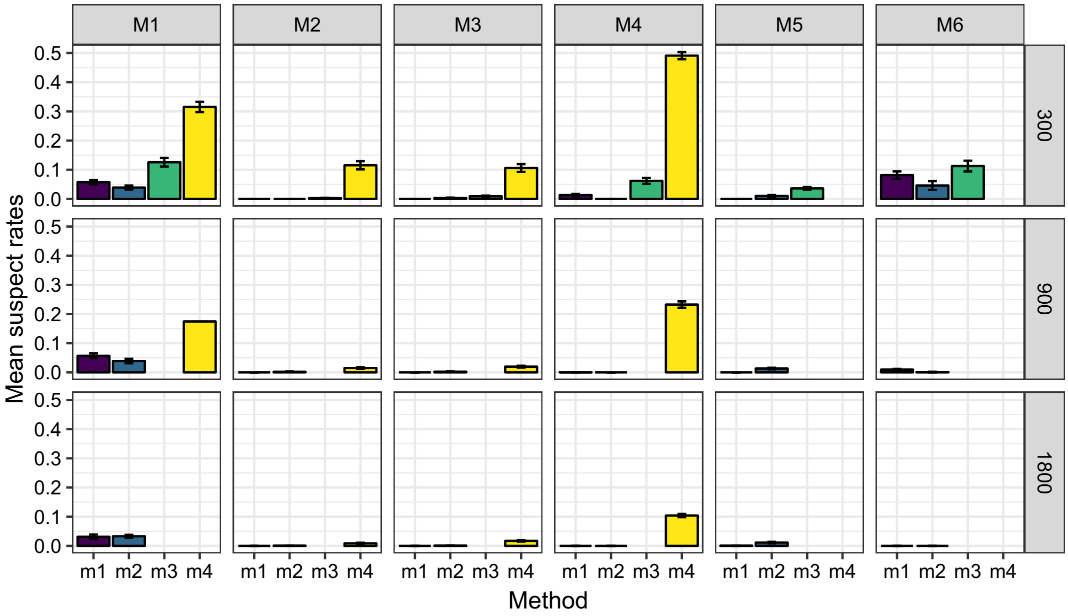

FIGURE 5 Mean suspect rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is represented by the whiskers. Each panel of the graph corresponds to some data model (defined by the column title) and some data size (defined by the row title) [Colour figure can be viewed at wileyonlinelibrary.com]

$\alpha = 0.05$, number of trees in a forest equal to 100, minimal amount of observations for splitting the node in a tree equal to 40 in the trees belonging to forests and equal to 60 in the PSICA tree.

Figure 8 shows that in four out of six nodes (Nodes 1, 2, 3, 5), supplementation options including early food supplementation had a larger probability of increasing HAZ at 54 m than the usual food supplementation, and this is in agreement with previous results of the trial,[29] Table A3. Similarly, our finding that in three out of six nodes (Nodes 4, 5, and 6), a supplementation including 30 mg of iron, and in two out of six nodes (Nodes 1 and 2) supplementation containing 60 mg of iron had higher HAZ compared to multiple micronutrient supplementation is also in agreement with previous results,[29] Table A3.

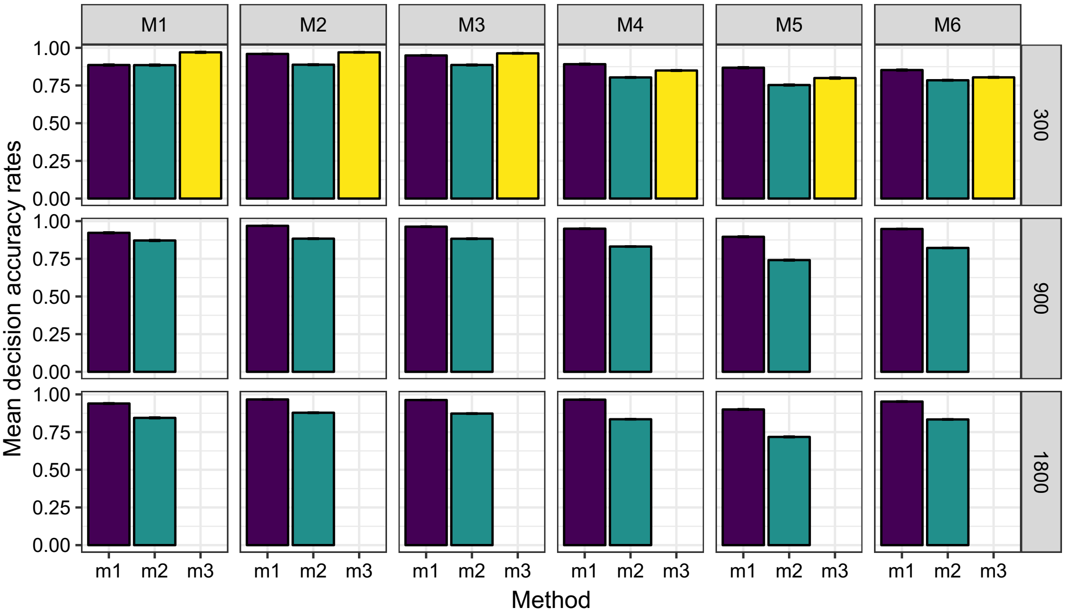

FIGURE 6 Mean decision rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is represented by the whiskers. Each panel of the graph corresponds to some data model (defined by the column title) and some data size (defined by the row title) [Colour figure can be viewed at wileyonlinelibrary.com]

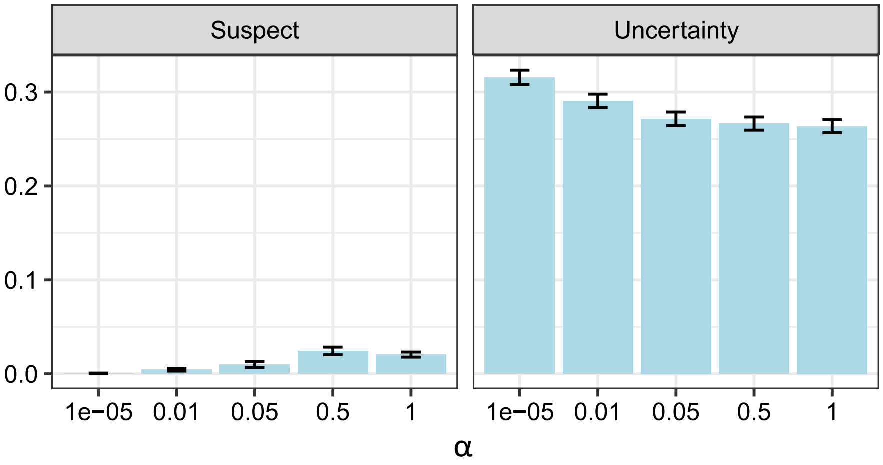

FIGURE 7 Mean uncertainty rates and mean suspect rates (over 400 experiments) for the data sets processed by m_1). Standard error of the mean is represented by the whiskers [Colour figure can be viewed at wileyonlinelibrary.com]

A result that has not been shown previously is that the optimal micronutrient supplementation varied with maternal height. Among the shortest women (Nodes 1 and 2), supplements containing 60 mg had the highest probability of a better HAZ in their offsprings. Among taller women (Nodes 3 to 6), supplements with a lower amount of iron (30 mg and MMS) had higher probabilities, and in three out of these four nodes, the optimal supplementation was 30 mg. While these differences in effects on young child height development have not previously been shown, they are biologically plausible in that shorter women are likely to have experienced more of nutrients deficiencies and thus larger nutrient requirements such as a larger dose of iron may be needed for optimal growth of their children. Maternal height has been shown to modify effect of micronutrient supplementation on other early life outcomes[30] and it is reasonable to believe that it will also modify other later outcomes. Similarly, indicators of socioeconomic situation such as maternal education have been shown to modify effect of micronutrient supplementation on early life outcome[30] and thus may also be of importance for young child height. The importance of iron for fetal, infant, and child growth has been shown in studies in low-income settings[31,32] and iron supplementation has been highlighted as a key intervention to improve maternal and children's health.[33]

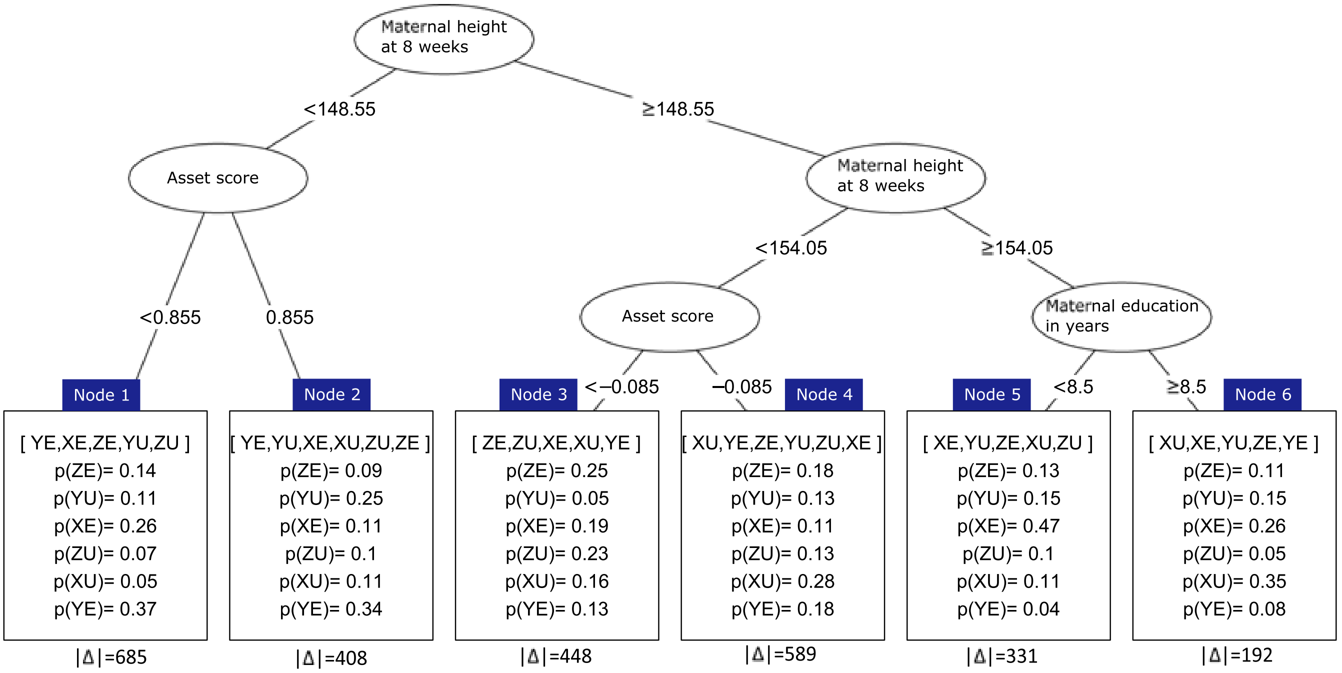

FIGURE 8 A PSICA tree showing subgroups and probabilities of various treatments for the MINIMat trial. Amounts of observations in the terminal nodes are represented by |Δ| [Colour figure can be viewed at wileyonlinelibrary.com]

4 | CONCLUSIONS AND DISCUSSION

In this work, we introduce PSICA trees. This is a novel method for subgroup identification in scenarios with categorical sets of treatments. Our numerical results illustrate that, with appropriate settings, PSICA trees provide high accuracies of prediction of the best treatments, and the method's uncertainty decreases with an increasing amount of data. At the same time, PSICA trees are easily interpretable and can therefore be used for policy making. The PSICA trees seem to be able to identify meaningful subgroups even when there are moderate mean effects from a lot of inputs, while in these cases, the QUINT method often fails to identify meaningful subgroups or it gives low accuracies. The PSICA trees are also able to handle cases when none of the treatments leads to a significantly better outcome than the other treatment: in this case, a noninformative tree (ie, in which all the treatments are declared to be best) can be returned.

It appears that PSICA trees providing the best accuracies are obtained when the amount of splitting variables in the corresponding random forest is equal to the total amount of inputs. There is also an indication of that bootstrapping random forests instead of using a bias-corrected infinitesimal jackknife might lead to lower uncertainties of the PSICA method. However, the price for this is greatly increased computational time. Some of the results also indicate that PSICA trees might not be very sensitive to the error's distribution.

The PSICA trees are computed by estimating probabilities and loss functions in a statistically motivated manner, which leads to high accuracies and low suspect rates in our simulation experiments. A real case study justifies the validity of our method because the information provided by the PSICA tree is also confirmed by previous medical studies.

The PSICA trees presented in this paper have some limitations. Firstly, the PSICA method was described for real-valued outcome variables only. We also assumed that $Y(x, \tau) = f(x, \tau) + \epsilon$, where ϵ is independent between different treatment options of the same patient, but this independence assumption might not hold in practice. However, since for a patient with some characteristics X_i, we only observe $Y(X_i, t_i)$ and never observe any other $Y(X_i, \tau)$ such that $\tau \neq t_i$, the observed data distribution will not depend on possible correlations $\mathrm{cor}(Y(X_i, t_i), Y(X_i, \tau))$. Since the observed data do not contain information on the magnitude of these correlations, we model ϵ as a term, which is independent between different treatment options of the same patient.

It was also assumed that randomized clinical trials data are used. Accordingly, a further research direction is to generalize the PSICA algorithm to categorical outcome scenarios and to investigate how it needs to be modified for non-randomized trials. Additionally, investigating the possibility of postpruning instead of prepruning might lead to a decrease in the suspect rates of the PSICA method.

ACKNOWLEDGEMENTS

The authors are grateful to the Swedish Research Council for providing funding for this work (project 2014-2161). We would also like to thank A. Rahman, S. el Arifeen, R. Naved, A. I. Khan, and L. Å. Persson for collecting and providing MINIMat data for our analysis.

DATA AVAILABILITY STATEMENT

The MINIMat data that support the findings of this study are available from the MINIMat group. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://doi.org/10.1186/ISRCTN16581394 with the permission of MINIMat group.

ORCID

Oleg Sysoev https://orcid.org/0000-0002-3092-4162

REFERENCES

1. Mayer C, Lipkovich I, Dmitrienko A. Survey results on industry practices and challenges in subgroup analysis in clinical trials. *Stat Biopharm Res*. 2015;7(4):272-282.
2. Jain KK. *Textbook of Personalized Medicine*. New York, NY: Springer; 2016.
3. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med*. 2010;363(4):301-304.
4. Lipkovich I, Dmitrienko A, D'Agostino Sr RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statist Med*. 2017;36(1):136-196.
5. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statist Med*. 2011;30(24):2867-2880.
6. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012;68(4):1010-1018.
7. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. *J Am Stat Assoc*. 2012;107(499):1106-1118.
8. Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. *Stat Methods Med Res*. 2013;22(5):493-504.
9. Foster JC, Taylor JMG, Kaciroti N, Nan B. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics*. 2014;16(2):368-382.
10. Zhang Y, Laber EB, Tsiatis A, Davidian M. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*. 2015;71(4):895-904.
11. Laber EB, Zhao YQ. Tree-based methods for individualized treatment regimes. *Biometrika*. 2015;102(3):501-514.
12. Fu H, Zhou J, Faries DE. Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statist Med*. 2016;35(19):3285-3302.
13. Doove LL, Dusseldorp E, van Deun K, van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Adv Data Anal Classif*. 2014;8(4):403-425.
14. Dusseldorp E, van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statist Med*. 2014;33(2):219-237.
15. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10:141-158.
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. *Statist Med*. 2011;30(21):2601-2621.
17. Thomas M, Bornkamp B, Seibold H. Subgroup identification in dose-finding trials via model-based recursive partitioning. *Statist Med*. 2018;37(10):1608-1624.
18. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. *J Comput Graph Stat*. 2008;17(2):492-514.
19. Loh W-Y, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Statist Med*. 2015;34(11):1818-1833.
20. Persson LA. Maternal and infant nutrition interventions in Matlab (MINIMat). https://doi.org/10.1186/ISRCTN16581394
21. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*. 2000;21(1):121-145.
22. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat*. 2006;15(3):651-674.
23. Wager S, Hastie T, Efron B. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J Mach Learn Res*. 2014;15(1):1625-1651.
24. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. Vol. 1. New York, NY: Springer; 2001. *Springer Series in Statistics*.
25. Breiman L. *Classification and Regression Trees*. Abingdon, UK: Routledge; 1984.

26. Sysoev O, Bartoszek K, Selling KE, Ekstrom L. psica: decision tree analysis for probabilistic subgroup identification with multiple treatments. R package version 1.0.0. 2018. https://CRAN.R-project.org/package=psica

27. Dusseldorp E, Doove L, van Mechelen I. Quint: an R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. *Behav Res Methods*. 2016;48(2):650-663.

28. Persson L, Arifeen S, Ekström EC, et al. Effects of prenatal micronutrient and early food supplementation on maternal hemoglobin, birth weight, and infant mortality among children in Bangladesh: the MINIMat randomized trial. *JAMA*. 2012;307(19):2050-2059.

29. Khan AI, Kabir I, Ekström E-C, et al. Effects of prenatal food and micronutrient supplementation on child growth from birth to 54 months of age: a randomized trial in Bangladesh. *Nutrition Journal*. 2011;10(1):134.

30. Smith ER, Shankar AH, Wu LS-F, et al. Modifiers of the effect of maternal multiple micronutrient supplementation on stillbirth, birth outcomes, and infant mortality: a meta-analysis of individual patient data from 17 randomised trials in low-income and middle-income countries. *Lancet Glob Health*. 2017;5(11):e1090-e1100.

31. Nisar YB, Dibley MJ, Aguayo VM. Iron-folic acid supplementation during pregnancy reduces the risk of stunting in children less than 2 years of age: a retrospective cohort study from Nepal. *Nutrients*. 2016;8(2):67.

32. Nguyen PH, Gonzalez-Casanova I, Young MF, et al. Preconception micronutrient supplementation with iron and folic acid compared with folic acid alone affects linear growth and fine motor development at 2 years of age: a randomized controlled trial in Vietnam. *J Nutr*. 2017;147(8):1593-1601.

33. Bhutta ZA, Ahmed T, Black RE, et al. What works? Interventions for maternal and child undernutrition and survival. *The Lancet*. 2008;371(9610):417-440.

How to cite this article: Sysoev O, Bartoszek K, Ekström E-C, Ekholm Selling K. PSICA: Decision trees for probabilistic subgroup identification with categorical treatments. *Statistics in Medicine*. 2019;38:4436–4452. https://doi.org/10.1002/sim.8308

APPENDIX

TABLE A1 Mean accuracy rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is specified in parentheses

n	model	m_1	m_2	m_3	m_4
300	1	1.00 (< 0.001)	1.00 (< 0.001)	1.00 (< 0.001)	0.96 (0.006)
900	1	1.00 (< 0.001)	1.00 (< 0.001)	–	1.00 (0.001)
1800	1	1.00 (< 0.001)	1.00 (< 0.001)	–	1.00 (< 0.001)
300	2	0.99 (0.001)	1.00 (< 0.001)	0.97 (0.002)	0.78 (0.013)
900	2	1.00 (< 0.001)	1.00 (< 0.001)	–	0.94 (0.005)
1800	2	1.00 (< 0.001)	1.00 (< 0.001)	–	0.95 (0.003)
300	3	0.99 (0.001)	1.00 (< 0.001)	0.97 (0.003)	0.73 (0.016)
900	3	1.00 (< 0.001)	1.00 (< 0.001)	–	0.93 (0.006)
1800	3	1.00 (< 0.001)	1.00 (< 0.001)	–	0.95 (0.004)
300	4	1.00 (< 0.001)	1.00 (< 0.001)	1.00 (0.001)	0.60 (0.009)
900	4	1.00 (0.001)	1.00 (< 0.001)	–	0.82 (0.014)
1800	4	0.99 (< 0.001)	1.00 (< 0.001)	–	0.98 (0.005)
300	5	0.98 (0.002)	0.99 (0.001)	0.97 (0.002)	–
900	5	0.98 (0.002)	1.00 (< 0.001)	–	–
1800	5	0.97 (0.002)	1.00 (< 0.001)	–	–
300	6	0.93 (0.007)	0.88 (0.016)	0.62 (0.019)	–
900	6	0.90 (0.003)	0.90 (0.013)	–	–
1800	6	0.88 (0.003)	0.91 (0.013)	–	–

TABLE A2 Mean uncertainty rates (over 200 experiments) for different data models (M1 to M6) processed by four methods ($m_1$to m_4). Standard error of the mean is specified in parentheses

n	Model	m_1	m_2	m_3	m_4
300	1	0.36 (0.008)	0.48 (0.014)	0.14 (0.012)	0.87 (0.020)
900	1	0.29 (0.009)	0.58 (0.014)	–	0.99 (0.003)
1800	1	0.26 (0.008)	0.71 (0.014)	–	1.00(< 0.001)
300	2	0.23 (0.009)	0.49 (0.003)	0.17 (0.010)	0.42 (0.014)
900	2	0.20 (0.007)	0.50 (0.001)	–	0.51 (0.008)
1800	2	0.25 (0.006)	0.50 (0.001)	–	0.50 (0.007)
300	3	0.29 (0.010)	0.50 (0.003)	0.20 (0.011)	0.44 (0.019)
900	3	0.24 (0.008)	0.50 (0.001)	–	0.51 (0.008)
1800	3	0.27 (0.007)	0.50 (0.001)	–	0.50 (0.007)
300	4	0.41 (0.004)	0.43 (0.002)	0.43 (0.002)	0.34 (0.011)
900	4	0.26 (0.009)	0.44 (0.001)	–	0.50 (0.017)
1800	4	0.14 (0.007)	0.44 (0.001)	–	0.66 (0.007)
300	5	0.45 (0.009)	0.67 (0.008)	0.52 (0.009)	–
900	5	0.39 (0.009)	0.77 (0.010)	–	–
1800	5	0.35 (0.008)	0.88 (0.010)	–	–
300	6	0.40 (0.004)	0.42 (0.002)	0.38 (0.007)	–
900	6	0.22 (0.008)	0.42 (0.001)	–	–
1800	6	0.09 (0.004)	0.42 (0.001)	–	–

TABLE A3 Mean suspect rates (over 200 experiments) for different data models (M1 to M6) processed by four methods (m_1 to m_4). Standard error of the mean is specified in parentheses

n	Model	m_1	m_2	m_3	m_4
300	1	0.06 (0.007)	0.04 (0.007)	0.13 (0.015)	0.32 (0.018)
900	1	0.06 (0.008)	0.04 (0.008)	–	0.17 (0.005)
1800	1	0.03 (0.008)	0.03 (0.005)	–	–
300	2	0.00 (<0.001)	0.00 (<0.001)	0.00 (0.001)	0.12 (0.014)
900	2	0.00 (<0.001)	0.00 (0.001)	–	0.02 (0.003)
1800	2	0.00 (<0.001)	0.00 (0.001)	–	0.01 (0.002)
300	3	0.00 (<0.001)	0.00 (0.001)	0.01 (0.002)	0.11 (0.013)
900	3	0.00 (<0.001)	0.00 (0.001)	–	0.02 (0.003)
1800	3	0.00 (<0.001)	0.00 (0.001)	–	0.02 (0.003)
300	4	0.01 (0.004)	0.00 (<0.001)	0.06 (0.010)	0.49 (0.012)
900	4	0.00 (0.001)	0.00 (<0.001)	–	0.23 (0.011)
1800	4	0.00 (<0.001)	0.00 (<0.001)	–	0.10 (0.006)
300	5	0.00 (<0.001)	0.01 (0.003)	0.04 (0.005)	–
900	5	0.00 (<0.001)	0.01 (0.003)	–	–
1800	5	0.00 (0.001)	0.01 (0.003)	–	–
300	6	0.08 (0.013)	0.05 (0.015)	0.11 (0.018)	–
900	6	0.01 (0.003)	0.00 (0.001)	–	–
1800	6	0.00 (<0.001)	0.00 (<0.001)	–	–

TABLE 4 Mean decision accuracy rates (over 200 experiments) for different data models (M1 to M6) processed by PSICA methods (m_1 to m_3). Standard error of the mean is specified in parentheses

n	Model	m_1	m_2	m_3
300	1	0.89 (0.003)	0.89 (0.003)	0.97 (0.003)
900	1	0.92 (0.003)	0.87 (0.003)	–
1800	1	0.94 (0.002)	0.84 (0.003)	–
300	2	0.96 (0.001)	0.89 (0.002)	0.97 (0.002)
900	2	0.97 (0.001)	0.88 (0.002)	–
1800	2	0.97 (0.001)	0.88 (0.002)	–
300	3	0.95 (0.002)	0.89 (0.003)	0.96 (0.002)
900	3	0.96 (0.001)	0.88 (0.002)	–
1800	3	0.96 (0.001)	0.87 (0.002)	–
300	4	0.89 (0.003)	0.80 (0.002)	0.85 (0.002)
900	4	0.95 (0.001)	0.83 (0.001)	–
1800	4	0.97 (0.001)	0.84 (0.001)	–
300	5	0.87 (0.003)	0.75 (0.003)	0.80 (0.004)
900	5	0.90 (0.002)	0.74 (0.002)	–
1800	5	0.90 (0.002)	0.72 (0.002)	–
300	6	0.85 (0.003)	0.78 (0.002)	0.80 (0.002)
900	6	0.95 (0.001)	0.82 (0.001)	–
1800	6	0.95 (0.001)	0.83 (0.001)	–

TABLE 5 Mean efficiency metrics (over 400 experiments) for different data models and data sizes processed by m_1 with different α values. Standard error of the mean is specified in parentheses

α	Accuracy	Decision accuracy	Suspect	Uncertainty
10^{-5}	0.982 (0.002)	0.932 (0.002)	0.000 (<0.001)	0.316 (0.008)
0.01	0.980 (0.002)	0.936 (0.002)	0.004 (0.001)	0.291 (0.007)
0.05	0.979 (0.002)	0.937 (0.002)	0.010 (0.003)	0.272 (0.007)
0.5	0.977 (0.002)	0.938 (0.002)	0.024 (0.004)	0.266 (0.007)
1	0.977 (0.003)	0.937 (0.002)	0.020 (0.003)	0.264 (0.007)