Web-based Supplementary Materials for Qualitative Interaction Trees: A tool to identify qualitative treatment-subgroup interactions

Elise Dusseldorp [1,2,*,†] and Iven Van Mechelen[2]

[1] Netherlands Organization for Applied Scientific Research (TNO), Statistics Group, Wassenaarseweg 56, Leiden, The Netherlands

[2] Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102 – bus 3713 , Leuven, Belgium

*Correspondence to: Elise Dusseldorp, TNO, Statistics Group, P.O. Box 2215, 2301 CE Leiden, The Netherlands

[†]E-mail: elise.dusseldorp@tno.nl

A. Choices to be made

A.1. *Which difference in treatment outcome?*

With regard to the Difference in treatment outcome component (see article, Section 2.3.1), the user needs to choose between differences in treatment means and treatment effect sizes. This choice is influenced by the specific application: e.g., if Y is measured on a scale, the values of which have a clear pragmatic meaning (such as the number of days spent in a hospital), crude differences in treatment means may be preferable. If the scale of Y is arbitrary, such as self-reported anxiety level measured through a Likert scale, treatment effect sizes may be a useful alternative. Also, in case overlap between the outcome distributions for the treatments ($T = 1$ and $T = 2$) is a key concern, the use of treatment effect size is preferable. We will clarify this with the following example: Suppose, the QUINT results in a tree with two leafs, the left node (R_1) is assigned to $\wp_1$ and the right node (R_2) is assigned to $\wp_2$. In R_1, the sample sizes of both treatments are 10, the mean for $T = 1$ is 5 with $s = 10$, and the mean for $T = 2$ is 4 with $s = 10$. Then, the mean difference is 1, and Cohen's d is 0.1. That would indicate that the overlap between the outcome distributions for the treatments is almost complete, and the choice between $T = 1$ and $T = 2$ trivial. In R_2, the sample sizes are both 100, the mean for $T = 1$ is 4 with $s = 1$, and the mean for $T = 2$ is 5 with $s = 1$. Then the mean difference (treatments reversed) is 1, and Cohen's d is 1.0. That would indicate almost complete separation between the outcome distributions. If you compared mean differences, you would say that the impact of treatment choice is the same in both groups, one favoring T = 1, the other favoring $T = 2$. However, if you compared effect sizes d, clearly the impact of treatment choice in R_2 is "sizeable" and that in R_1 is trivial. In conclusion, making decisions on the basis of an effect size, like Cohen's d, is often better.

A.2. *Choice of the weights*

The weights (w_1, and w_2) of the partitioning criterion (see article, Section 2.3.3) are pre-specified. The values are chosen in such a way that the Difference in treatment outcome component and the Cardinality component are put on comparable measurement scales. This is realized by setting the realistic maximum of both components after weighting about equal. The largest possible difference in treatment means within a node is the range of Y; instead of the range of Y, we take a more robust value: the interquartile range (IQR) of Y. For treatment effect size (d), a realistic maximum is $d = 3$. As a consequence, as value for w_1 we propose $1/\log(1+IQR(Y))$ in case the Difference in treatment outcome is measured in terms of difference in treatment means, and 1/log(1+3) in case it is measured in terms of treatment

effect size. With regard to the Cardinality component, the maximum value is achieved when half of the subjects are assigned to $\wp_1$ and half of the subjects to $\wp_2$. Therefore, as value for w_2 we propose 1 / log(0.50*N*). One may note that the proposed values result in a realistic maximum of 2 for each component, and thus a value of 4 for the criterion *C*.

A.3. *Default values for stopping criteria*

For $L_{\text{upperlimit}}$, that is, the a priori fixed maximum number of leaves, the default was set at 10. In our analyses with QUINT, the value of $L_{\max}$ (i.e., the maximum possible number of leaves, because the maximum value of the partitioning criterion *C* was reached) never exceeded 10.

For the *qualitative interaction condition,* a value of $d_{\min}$ (i.e., the minimum absolute standardized mean difference in treatment outcome in each of the two leaves) has to be chosen a priori. The default value of $d_{\min}$ was set at 0.30, which was based on the results of the Simulation study (see article, Section 3.5).

For the *minimal sample size per treatment condition*, we chose as default values 10% of the subjects in each treatment group: $0.10n_A$ and $0.10n_B$, where n_A and n_B denote the sample sizes of treatment A and B.

B. Pruning algorithm of QUINT

The pruning procedure starts by estimating trees for a range of tree sizes *L* (= number of leaves), $L = 2,\ldots,L_{\max}$ (=maximum possible number of leaves), for the total sample of size *N*. For each value of *L,* the procedure then proceeds as follows:

- First, the value of the partitioning criterion *C* (denoted by C_L^{app}) implied by the corresponding QUINT model for the total sample is calculated.
- Then *B* bootstrap samples of size *N* are drawn (a good choice is *B* = 25, see [26]).
- Subsequently, for each bootstrap sample *b*, a tree of size *L* is grown, and the value of the partitioning criterion *C* is computed (denoted by $C_{b,L}^{boot}$).
- Next, the bootstrap tree is "frozen"(i.e. the splitting variables, the split points and the assignment to the partition classes are fixed), and the performance of the tree on the original total sample is evaluated in terms of the corresponding value of the partitioning criterion (denoted by $C_{b,L}^{orig}$).
- The overoptimism in the fit of a tree of size *L* is then estimated by calculating for each bootstrap sample *b* :

  $O_{b,L} = C_{b,L}^{boot} - C_{b,L}^{orig}$,

and then averaging these values across all bootstrap samples b (b = 1,…, B):

$$\bar{O}_L = \frac{1}{B}\sum_{b=1}^{B} O_{b,L},$$

with corresponding standard error:

$$SE_L = \sqrt{\frac{\sum_{b=1}^{B}(O_{b,L} - \bar{O}_L)^2}{B-1}} / \sqrt{B}.$$

- Finally, the bias-corrected performance of the tree of size L is computed as:

  $C_L^{bc} = C_L^{app} - \bar{O}_L$.

After carrying out this whole procedure for all values of L, with L = 2,…,L_{max}, the optimal tree size is selected using the following so-called one standard error rule [27]: Let L^* denote the tree with the highest bias-corrected performance value C_L^{bc}; then, the size of the optimally pruned tree (L^{**}) corresponds to the minimum value of L which is such that:

$$C_{L^{**}}^{bc} \geq C_{L^*}^{bc} - SE_{L^*}^{bc}.$$

In the *R*-package "quint", a QUINT analysis incorporates automatically the bias-corrected bootstrap procedure. The default value of B is set at 25 (recommended by [26]). A separate function is available for selecting the optimal tree size by applying the one standard error rule.

C. Validation procedures for the effect sizes

To get insight into the generalizability of a QUINT solution, we recommend to perform a validation procedure. We propose two validation procedures. The first one, which provides estimates of effect sizes and treatment outcomes for future observations, is clearly the preferred one in case of RCTs with many observations. For smaller data sets ($N \leq 400$), we propose a second procedure based on a bootstrap methodology.

C.1. *Validation procedure for relatively large data sets (N > 400)*

Before starting the analysis, the data are randomly divided into a training data set (75% of the observations) and a test data set (25% of the observations). Perform a full QUINT analysis using the training set: Grow a large tree, and apply the pruning procedure. Next, freeze the pruned tree of the training set, and pass the test set through the tree. Use the resulting differences in treatment outcome in the leafs of the tree as the estimates of the population values.

C.2. *Validation procedure for small data sets*

Start with performing a full QUINT analysis using the total observed data set: Grow a large tree, and apply the pruning procedure to determine the optimal number of leaves (L_{optimal}). Next, draw B bootstrap samples from the observed data set, and perform a QUINT analysis for each bootstrap sample with $L_{\text{upperlimit}}$ equal to L_{optimal}. Compute for each bootstrap sample, the difference between the two leafs with the largest positive effect size ($d_{b,\max}^{boot}$) and largest negative effect size ((i.e. $d_{b,\min}^{boot}$), that is, the range of the effect sizes in the bootstrap sample. Next, the tree of each bootstrap sample is "frozen", and the original data are passed through the tree. Compute the difference in the effect sizes of the same "extreme" nodes: $d_{b,\max}^{orig} - d_{b,\min}^{orig}$. Next, average these values across all bootstrap sample b (b = 1,…, B), to obtain an estimate of selection bias (i.e., the mean optimism in the range of the effect sizes):

$$\overline{O}_{range} = \frac{1}{B}\sum_{b=1}^{B}(d_{b,\max}^{boot} - d_{b,\min}^{boot}) - \frac{1}{B}\sum_{b=1}^{B}(d_{b,\max}^{orig} - d_{b,\min}^{orig}).$$

Table 1. Goodness-of-recovery of the true tree structure by QUINT. Cells display the proportion of QUINT solutions in the true tree size that yield the true splitting variables (left) and the conditional proportions (for the solutions with the true splitting variables) of the true split points ± 5 (right). Results are displayed for the situations with correlation between covariates ($\rho = .20$).

			True splitting variables *Model*				True split point ± 5 *Model*			
J	*DT*	*N*	A	B	C	D	A	B	C	D
5	M	200	.76	.26	.10	.09	.66	.77	.50	.67
		300	.89	.41	.13	.06	.82	.76	.62	.33
		400	.95	.45	.28	.26	.91	.80	.50	.54
		500	.97	.70	.32	.20	.93	.76	.69	.65
		1000	1.00	.90	.65	.56	1.00	.96	.91	.86
	L	200	.99	.82	.37	.33	.98	.91	.65	.88
		300	1.00	.97	.65	.49	.99	.92	.75	.84
		400	1.00	1.00	.81	.66	1.00	.98	.83	.83
		500	1.00	.99	.94	.68	1.00	.99	.85	.91
		1000	1.00	1.00	1.00	.91	1.00	1.00	.97	.90
	XL	200	1.00	1.00	.84	.81	1.00	.95	.85	.91
		300	1.00	1.00	.95	.79	1.00	.95	.84	.87
		400	1.00	1.00	.96	.83	1.00	.97	.94	.88
		500	1.00	1.00	1.00	.90	1.00	1.00	.93	.89
		1000	1.00	1.00	1.00	.96	1.00	1.00	.99	.95
10	M	200	.63	.13	.02	.00	.60	.77	.00	.00
		300	.79	.30	.05	.05	.78	.83	.60	.80
		400	.91	.33	.08	.06	.88	.82	.75	1.00
		500	.99	.46	.22	.14	.97	.91	.64	.71
		1000	1.00	.86	.61	.37	1.00	.95	.93	.92
	L	200	.98	.69	.25	.17	.97	.96	.76	.82
		300	1.00	.93	.48	.38	1.00	.97	.94	.89
		400	1.00	.97	.76	.55	1.00	.99	.82	.85
		500	1.00	.99	.82	.60	1.00	.93	.87	.85
		1000	1.00	1.00	.99	.78	1.00	1.00	.96	.94
	XL	200	1.00	.99	.88	.55	1.00	.98	.78	.89
		300	1.00	1.00	.98	.82	1.00	.98	.92	.93
		400	1.00	1.00	.99	.93	1.00	.99	.91	.91
		500	1.00	1.00	1.00	.90	1.00	.98	.90	.94
		1000	1.00	1.00	1.00	.92	1.00	1.00	.97	.92
20	M	200	.58	.07	.00	.00	.55	.57	.00	.00
		300	.74	.19	.00	.02	.72	.84	.00	1.00
		400	.84	.20	.07	.03	.79	.85	.57	.67
		500	.97	.46	.08	.06	.92	.85	1.00	.83
		1000	1.00	.80	.51	.18	.99	.98	.86	1.00
	L	200	.98	.67	.13	.11	.98	.81	.31	.91
		300	.99	.85	.43	.38	.98	.99	.77	.84
		400	1.00	.94	.71	.46	1.00	.97	.90	.87
		500	1.00	.97	.86	.53	1.00	.96	.87	.85
		1000	1.00	1.00	.99	.85	1.00	1.00	.96	.91
	XL	200	1.00	.98	.76	.57	1.00	.92	.74	.88
		300	1.00	1.00	.98	.80	1.00	.97	.84	.94
		400	1.00	1.00	.99	.76	1.00	.99	.94	.95
		500	1.00	1.00	1.00	.90	1.00	.96	.94	.89
		1000	1.00	1.00	1.00	.99	1.00	1.00	.99	.95

Note. J = total number of covariates; *DT* = Difference in treatment outcome. M=Medium; L=Moderately large; XL=Very large; *N*= Sample size.

Table 2. Goodness-of-recovery of the true assignment to the partition classes. Cells display the mean agreement (mean Cohen's κ) between the assignments of the QUINT solutions and the true assignments.

		Cohen's κ			
		Model			
DT	*N*	A	B	C	D
M	200	.58	.19	.13	.09
	300	.74	.28	.20	.15
	400	.84	.36	.24	.18
	500	.91	.43	.32	.25
	1000	.97	.67	.61	.44
L	200	.95	.62	.41	.36
	300	.98	.80	.61	.53
	400	.99	.88	.74	.64
	500	.99	.89	.82	.71
	1000	1.00	.97	.95	.87
XL	200	.99	.92	.83	.72
	300	1.00	.95	.92	.82
	400	1.00	.96	.93	.86
	500	1.00	.97	.95	.89
	1000	1.00	.98	.97	.94

Note. Results are presented separately for simulated data with varying true size of the difference in treatment outcome (*DT*) (i.e., medium [M], moderately large [L], and very large [XL]), and with varying sample size *N*. Results have been averaged across the levels of the factors number of covariates (5, 10, or 20), and intercorrelation between covariates ($\rho = 0$, or $\rho = 0.20$).

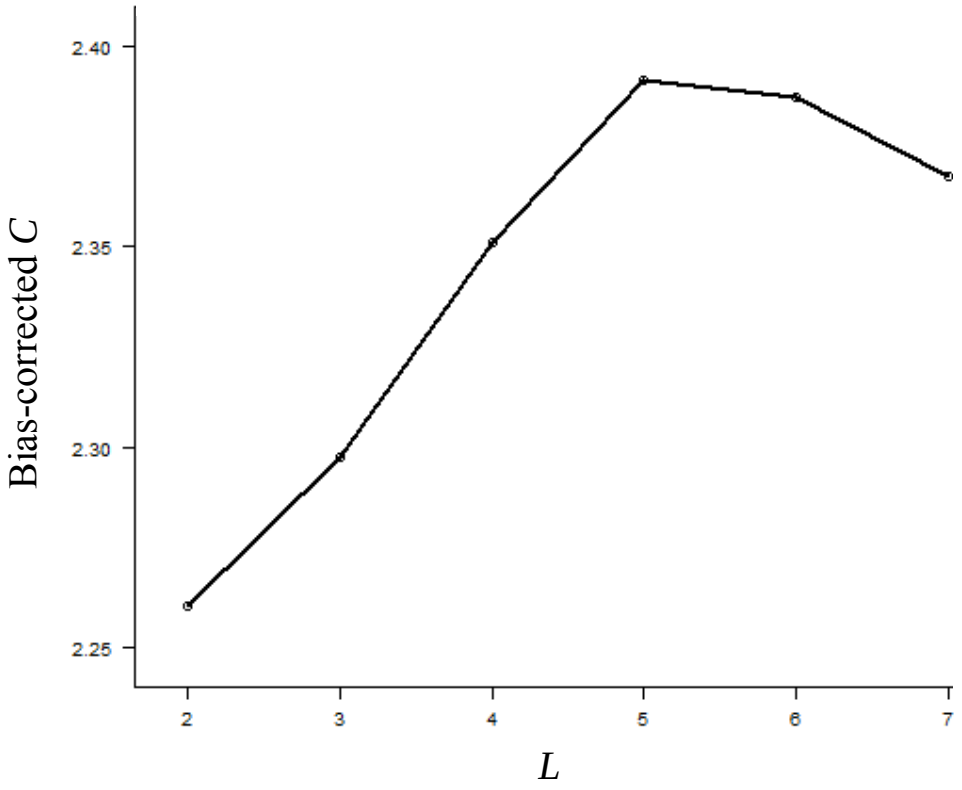

Figure 1. Bias-corrected criterion value of the partitioning criterion (*C*) plotted against the number of leaves *L* of the QUINT solution for the data from the Breast Cancer Recovery Project [23]. The number of bootstraps (*B*) was set at 200.