Psychotherapy Research

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: http://www.tandfonline.com/loi/tpsr20

# QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials

Lisa L. Doove, Katrijn Van Deun, Elise Dusseldorp & Iven Van Mechelen

**To cite this article:** Lisa L. Doove, Katrijn Van Deun, Elise Dusseldorp & Iven Van Mechelen (2015): QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials, Psychotherapy Research, DOI: 10.1080/10503307.2015.1062934

**To link to this article:** http://dx.doi.org/10.1080/10503307.2015.1062934

Published online: 14 Jul 2015.

*Psychotherapy Research*, 2015
http://dx.doi.org/10.1080/10503307.2015.1062934

METHOD PAPER

# QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials

LISA L. DOOVE[1], KATRIJN VAN DEUN[1,2], ELISE DUSSELDORP[1,3], & IVEN VAN MECHELEN[1]

[1]*Department of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Leuven, Belgium;* [2]*Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands &* [3]*Mathematical Institute, Leiden University, Leiden, The Netherlands*

*(Received 31 July 2014; revised 8 April 2015; accepted 9 June 2015)*

**Abstract**

**Objective**: The detection of subgroups involved in qualitative treatment–subgroup interactions (i.e., for one subgroup of clients treatment A outperforms treatment B, whereas for another the reverse holds true) is crucial for personalized health. In typical Randomized Controlled Trials (RCTs), the combination of a lack of a priori hypotheses and a large number of possible moderators leaves current methods insufficient to detect subgroups involved in such interactions. A recently developed method, QUalitative INteraction Trees (QUINT), offers a solution. However, the paper in which QUINT has been introduced is not easily accessible for non-methodologists. In this paper, we want to review the conceptual basis of QUINT in a nontechnical way, and illustrate its relevance for psychological applications. **Method**: We present a concise introduction into QUINT along with a summary of available evidence on its performance. Subsequently, we subject RCT data on the effect of motivational interviewing in a treatment for substance abuse disorders to a reanalysis with QUINT. As outcome variables, we focus on measures of retention and substance use. **Results**: A qualitative treatment–subgroup interaction is found for retention. By contrast, no qualitative interaction is detected for substance use. **Conclusions**: QUINT may lead to insightful and well-interpretable results with straightforward implications for personalized treatment assignment.

**Keywords:** qualitative interaction; subgroup analysis; treatment efficacy; QUINT

For many psychological problems, multiple treatment alternatives are available (e.g., De Jong et al., 2014; Gullestad, Johansen, Høglend, Karterud, & Wilberg, 2013; Tasca et al., 2006). As an example, Tasca et al. (2006) report two forms of group treatment (cognitive-behavioral therapy, psychodynamic interpersonal psychotherapy) for persons with binge eating disorder. A standard research question in such cases pertains to comparative treatment effectiveness, that is, to a comparison of the effect of the different treatments. A typical setting for the study of this type of research questions is that of randomized controlled trials (RCTs), in which the clients under study are randomly assigned to the alternative treatment conditions, and in which they are measured in terms of a set of pre-treatment characteristics in addition to (at least) one outcome variable. Analyses of such RCT data may reveal which treatment alternative is universally best (i.e., yields the best mean outcome), a result that may have direct implications for treatment assignment (i.e., assign every future patient to the universally best treatment).

Beyond an assignment to universally best treatments, nowadays much importance is attached to personalized health, that is, to the fact that persons should receive the treatment that is optimal given their individual characteristics (Tunis, Benner, & McClellan, 2010). Personalized health is rooted in the idea that relative treatment effectiveness may vary over subgroups of clients (defined in terms of

Correspondence concerning this article should be addressed to Lisa L. Doove, Department of Psychology and Educational Sciences, Katholieke Universiteit Leuven, Tiensestraat 102 – bus 3713, Leuven, Belgium. Email: lisa.doove@ppw.kuleuven.be

person pre-treatment characteristics, which are also called moderator variables). Such a situation is illustrated in Figure 1(a) and (b). Formally speaking, this phenomenon is referred to as treatment–subgroup interactions (Rothwell, 2005). One type of such interactions is especially relevant for treatment assignment, namely so-called disordinal or qualitative treatment–subgroup interactions (Byar, 1985) (note that the meaning of the term "qualitative" as used in the context of treatment–subgroup interactions should not be confused with the meaning of the same term in the context of qualitative research methods). Given two treatment alternatives A and B, qualitative treatment–subgroups interactions imply that the sign of the difference between the treatment alternatives in effectiveness is not the same for all subgroups of persons (Figure 1(b)). Or, stated in other words, qualitative treatment–subgroup interactions imply that for some subgroups treatment alternative A outperforms alternative B, whereas for other subgroups it is the other way around. Quantitative treatment–subgroup interactions, on the other hand, imply that the difference in treatment effectiveness has the same sign in all subgroups but that the size of the difference in treatment effectiveness is different (Figure 1(a)). Or, stated in other words, quantitative treatment–subgroup interactions imply that for all subgroups one treatment alternative (e.g., A) is more effective than or equally effective as the other, but that the magnitude of the difference in effectiveness between the two treatment alternatives varies across subgroups. As an example of a qualitative treatment–subgroup interaction, in the case of supportive versus interpretative therapy for depression, Ogrodniczuk, Piper, Joyce, and McCallum (2001) found that for males interpretative therapy outperforms supportive therapy, whereas for females the reverse holds true. As a second example, in the case of cognitive-behavioral therapy versus psychodynamic interpersonal psychotherapy for binge eating disorder, Tasca et al. (2006) found that for people with lower need for approval cognitive-behavioral therapy outperforms psychodynamic interpersonal psychotherapy, whereas for people with higher need for approval the reverse holds true. Qualitative treatment–subgroup interactions have a long history in psychology, going back to the seminal work of Cronbach (1957) on aptitude–treatment interactions. Importantly, despite the fact that in psychotherapy research such interactions may be rarer than their quantitative counterparts (Peto, 1995), they are of utmost clinical importance for personalized health (Byar, 1985).

Claims on qualitative treatment–subgroup interactions, as the cornerstone of personalized treatment assignment, should rely on strong empirical evidence, especially since evidence-based practice has become the gold standard in clinical psychology (Kent, Rothwell, Ioannidis, Altman, & Hayward, 2010). The detection of empirically sound interactions, however, implies a major methodological bottleneck. Earlier work on this detection primarily concerned two families of methods. The first and larger family pertains to situations in which clear a priori hypotheses exist about which subgroups of clients are involved in the interactions, or situations that involve a small number of potential moderator variables only. Examples include factorial analyses of variance (ANOVA), with one factor pertaining to treatment methods and another one to subgroups (Shaffer, 1991), and regression analyses with suitable interaction terms being included in the regression model (see, e.g., Dixon & Simon, 1991; Hayward, Kent, Vijan, & Hofer, 2006). Methods of the second and smaller family do not require a priori hypotheses or a limited number of potential moderator variables. Rather, they induce subgroups involved in treatment–subgroup interactions during the actual data analysis. In particular, they do so via recursive partitioning, in which the total group of persons is repeatedly split into child subgroups that vary in terms of relative treatment effectiveness (for a review, see Doove, Dusseldorp, Van Deun, & Van Mechelen, 2014).

However, the families of methods outlined above imply two problems. The first of these pertains to the methods being unsuitable to detect qualitative treatment–subgroup interactions in regular RCTs. Indeed, in many RCTs, no comprehensive a priori hypotheses are available on the subgroups involved in treatment–subgroup interactions (Bala et al., 2013; Boonacker, Hoes, Van Liere-Visser, Schilder, & Rovers, 2011), and a large number of potential moderators are available in the data. Such situations are prohibitive for ANOVA and regression-type approaches. Recursive partitioning methods can deal with such situations, yet they do so with a focus on treatment–subgroup interactions in general rather than on the qualitative interactions that are of particular relevance for personalized treatment assignment.

The second problem pertains to the risk of inferential errors. On the one hand, these include Type II errors, which reflect a possible lack of power to detect true interactions (e.g., Pocock, Assmann, Enos, & Kasten, 2002). Such a detection generally requires larger samples than the detection of main effects (Lee, Lei, & Brody, 2015), indeed, and perhaps considerably larger than those enrolled in a number of traditional clinical trials. On the other hand, and even more importantly, one should also

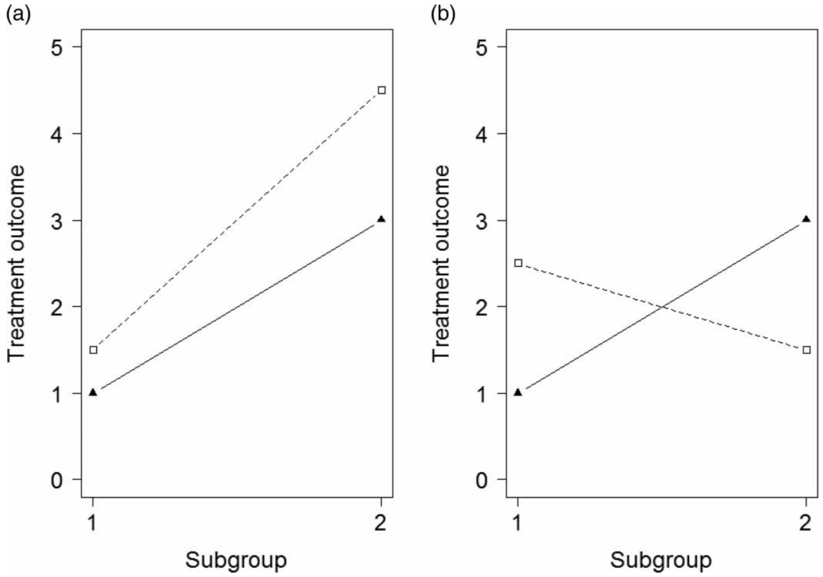

Figure 1. Examples of treatment-subgroup interactions: (a) a quantitative treatment-subgroup interaction, where the difference between the two treatment alternatives in treatment effectiveness has the same sign in both subgroups but the size of the effect differs, and (b) a qualitative treatment-subgroup interaction, where the difference between the two treatment alternatives in treatment effectiveness has a different sign in Subgroup 1 than in Subgroup 2. The subgroups may be defined in terms of several person pre-treatment characteristics. (▲, treatment A; □, treatment B).

beware of Type I errors, that is, erroneous claims about the occurrence of apparent interactions that cannot be replicated in follow-up studies (Dixon & Simon, 1991; Pocock et al., 2002; Rothwell, 2005; Wang, Lagakos, Ware, Hunter, & Drazen, 2007). Because of a presumed relatively high risk of Type I errors, a number of critics remembered subgroup analyses with the pet name of "computerized data dredging" (Feinstein, 1998; Rothwell, 2005). Beyond this, some critics even went so far as to dismiss all post hoc tests and to allow only for subgroups that are specified on a priori grounds. Obviously, however, such a recommendation passes over the fact that often no good hypotheses are available on the subgroups involved in treatment–subgroup interactions, and precludes exploratory approaches that allow RCT data to speak for themselves.

Recently, Dusseldorp and Van Mechelen (2014) developed a new method, called QUalitative INteraction Trees (QUINT), that addresses the two problems outlined above. This method, which is of a recursive partitioning type, allows subgroups involved in treatment–subgroup interactions to be identified in situations where a large number of moderators are available in the data, without comprehensive a priori hypotheses on such subgroups, and with an exclusive focus on qualitative treatment–subgroup interactions. QUINT is an exploratory approach that can be useful in situations where a clinician: (i) has no hypotheses on subgroups involved in qualitative treatment–subgroup interactions, (ii) has incomplete hypotheses on such subgroups in the sense that the exact nature of the interplay between different possible moderators is not known or that the clinician does not have precise hypotheses on the cut-off scores that define the subgroups, or (iii) has clear a priori hypothesis on some subgroups involved in treatment–subgroup interactions, but additionally also wants to explore the data somewhat further. The development of QUINT included an explicit account of the problem of inferential errors, in terms of an extensive simulation study that led to a number of strategies and recommendations to control for this problem. Unfortunately, however, the paper in which the QUINT methodology has been introduced is not easily accessible for non-methodologists.

In the present paper we will present a nontechnical review of the conceptual basis of QUINT and show its significance for psychological applications. To this end we will subject data from an RCT on drug abuse treatments to a reanalysis with QUINT. This will make it possible to arrive at a more pronounced picture of the information on treatment effectiveness embedded in the data.

The remainder of this paper is organized as follows. First, we will review the QUINT methodology. Second, we will introduce the RCT data that we will re-analyze using QUINT. Third, we will present and discuss the results of the reanalysis. Concluding remarks will be given in a final section.

## QUINT

Suppose a group of clients randomly assigned to one out of two treatments A and B. Before the treatment a group of categorical and/or continuous background characteristics of the clients is measured (e.g., addiction severity, primary drug used), and after the treatment one primary continuous outcome variable (which, optionally, can be a pre-post difference score). The goal of QUINT is to find the best partition of the total group of clients on the basis of the background characteristics into two or three mutually exclusive subgroups that are characterized as follows: In the first subgroup ($\wp_1$), the clients assigned to treatment A show a clearly better outcome than the clients assigned to B; in the second subgroup ($\wp_2$), the reverse is true; in the third (optional) subgroup ($\wp_3$), the clients assigned to A show more or less the same outcome as the clients assigned to B (Dusseldorp & Van Mechelen, 2014). The subgroups may comprise one or several types of clients as defined by different (combinations of) background characteristics. Note, however, that the result of a QUINT analysis may also be that the total group of clients is not partitioned, that is, that no subgroups involved in a qualitative treatment–subgroup interaction can be identified.

As mentioned earlier, QUINT is of a recursive partitioning type, which implies that the total group of clients is repeatedly subdivided on the basis of binary splits of the background characteristics into child subgroups. Classification And Regression Trees (CART) may be the best known instance of recursive partitioning analysis, first introduced by Breiman, Friedman, Olshen, and Stone (1984). Both methods, however, differ in the type of partitioning criterion that is used. In CART, a grouping variable such as the preferred treatment for every client in the data should be known in advance, with in each step the child subgroups being as homogeneous as possible with regard to that grouping variable. In contrast, QUINT does not need external grouping information to partition the total group of clients into subgroups that vary in terms of relative treatment effectiveness.

QUINT is looking for an optimal partition of the total group of clients so that the qualitative treatment-subgroup interaction that is related to that partition has the largest possible practical significance. To achieve this, two conditions with regard to the subgroups $\wp_1$ and $\wp_2$ need to be satisfied: (i) In both subgroups the difference in outcome between the treatments A and B should be large and (ii) each of the two subgroups should comprise many clients. QUINT uses a weighted compound criterion that implies that these two conditions are optimized simultaneously. The difference in outcome between treatments A and B included in condition (i) can be formalized in terms of either a difference in treatment means or a treatment effect size (Cohen's *d*; Cohen, 1988), with a difference in treatment means being preferable if the values of the outcome variable have a clear pragmatic meaning, whereas in other cases treatment effect sizes may be more useful.

To optimize this criterion, QUINT uses a stepwise tree building algorithm. This algorithm sequentially splits the total group of clients into subgroups, with the resulting series of splits being representable by a tree structure like Figure 2 (which we will further discuss in the Results Section). Starting with the total group of clients in the so-called root node, each background variable is considered as a candidate splitting variable to divide this group into two child nodes. For each of the candidate splitting variables, all possible split points and corresponding assignments of the child nodes to subgroups $\wp_1$ and $\wp_2$ are evaluated; subsequently, the split point and assignment of the child nodes are chosen that maximize the QUINT criterion. Lastly, across all candidate splitting variables, the variable (along with its maximizing split point and assignment of child nodes to the subgroups) is selected that attains the highest value of the QUINT criterion. After this first split, the stepwise binary splitting procedure is continued. In each step, all end nodes (leaves) of the current tree then become candidate parent nodes. For each candidate parent node, the split (i.e., splitting variable, split point, and assignment of all end nodes or leaves of the tree to subgroups $\wp_2$ $\wp_2$, and $\wp_3$) is selected that maximizes the QUINT criterion. The QUINT criterion values are subsequently compared across all candidate parent nodes and the node which implies the highest criterion value then is subdivided according to its optimal split. Note that from the second split on, leaves may be assigned to $\wp_1$, $\wp_2$, and $\wp_3$ (instead of to $\wp_1$ and $\wp_2$ only as after the first split), and that after each split all leaves are allowed to be re-assigned to the three subgroups.

The QUINT procedure uses three types of criteria to stop the tree building process: Firstly, in the split of the root node it tests the presence of a qualitative interaction on the basis of a so-called qualitative interaction condition, which reads that in each of the two leaves, the absolute value of the treatment effect size exceeds a critical minimum value ($d_{\min}$). Note that this critical minimum value is always formulated in terms of an effect size, irrespective of whether the difference in means or the effect size is used in condition (i) of the weighted compound criterion that is optimized by QUINT, as the critical minimum value has to hold irrespective of the scale

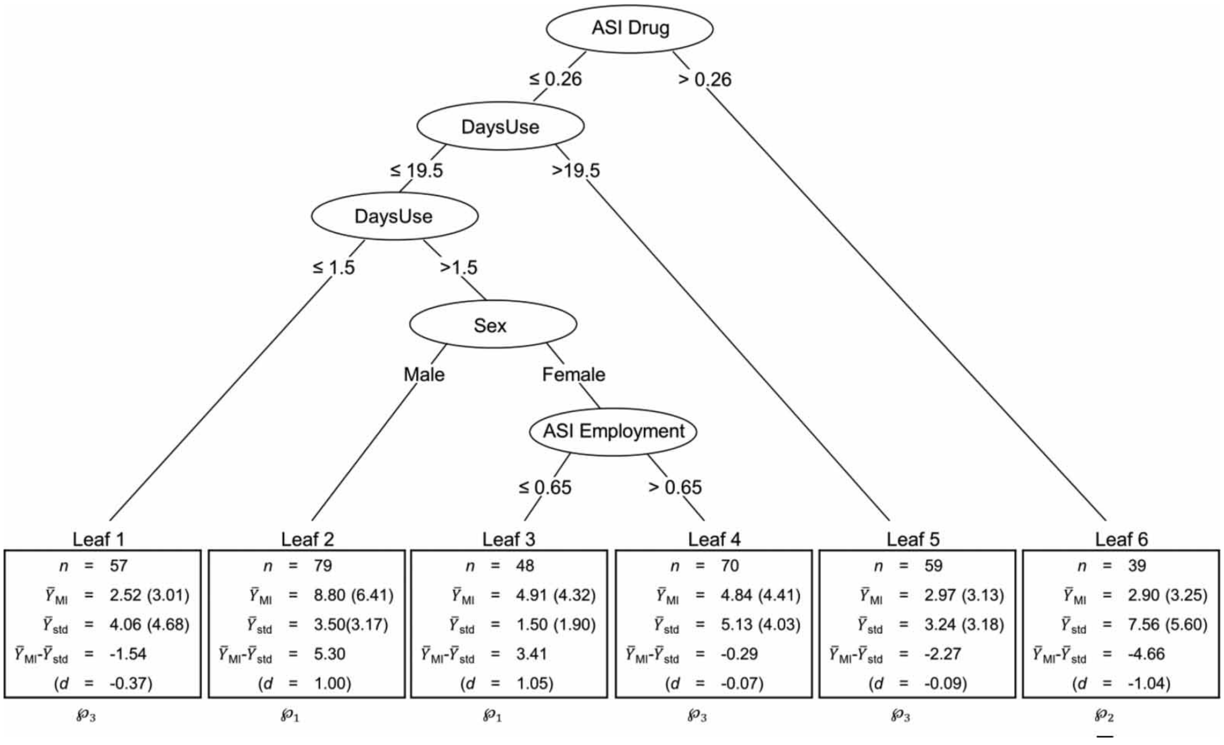
ASI Drug
≤ 0.26
> 0.26
DaysUse
≤ 19.5
>19.5
DaysUse
≤ 1.5
>1.5
Sex
Male
Female
ASI Employment
≤ 0.65
> 0.65
Leaf 1
n = 57
$\bar{Y}_{MI}$ = 2.52 (3.01)
$\bar{Y}_{std}$ = 4.06 (4.68)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = -1.54
(d = -0.37)
$\wp_3$
Leaf 2
n = 79
$\bar{Y}_{MI}$ = 8.80 (6.41)
$\bar{Y}_{std}$ = 3.50(3.17)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = 5.30
(d = 1.00)
$\wp_1$
Leaf 3
n = 48
$\bar{Y}_{MI}$ = 4.91 (4.32)
$\bar{Y}_{std}$ = 1.50 (1.90)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = 3.41
(d = 1.05)
$\wp_1$
Leaf 4
n = 70
$\bar{Y}_{MI}$ = 4.84 (4.41)
$\bar{Y}_{std}$ = 5.13 (4.03)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = -0.29
(d = -0.07)
$\wp_3$
Leaf 5
n = 59
$\bar{Y}_{MI}$ = 2.97 (3.13)
$\bar{Y}_{std}$ = 3.24 (3.18)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = -2.27
(d = -0.09)
$\wp_3$
Leaf 6
n = 39
$\bar{Y}_{MI}$ = 2.90 (3.25)
$\bar{Y}_{std}$ = 7.56 (5.60)
$\bar{Y}_{MI}$-$\bar{Y}_{std}$ = -4.66
(d = -1.04)
$\wp_2$

Figure 2. Result of the application of QUINT to the Clinical Trials Network data with the outcome variable “Number of sessions in 28 days after treatment assignment.” The resulting tree has six leaves that are represented by rectangles containing the sample size, the outcome means (and standard deviations) for the motivational interviewing and standard treatment groups ($\overline{Y}_{\mathrm{MI}}$ and $\overline{Y}_{\mathrm{std}}$), the differences in means ($\overline{Y}_{\mathrm{MI}} - \overline{Y}_{\mathrm{std}}$), and the corresponding effect size (Cohen’s *d*) (ASI = Addiction Severity Index).

of the outcome variable under study. If QUINT does detect a qualitative treatment–subgroup interaction, the root node is split and the stepwise binary splitting procedure is continued. Secondly, the algorithm stops if a split can no longer be found that implies a higher criterion value than in the previous step. Thirdly, QUINT takes into account some additional stopping criteria including a maximum value for the number of leaves, and the fact that each leaf should contain a minimum number of clients assigned to treatments A and B.

The tree growing procedure may result in a large tree with a high criterion value for the data at hand, that will not be replicable with future data. QUINT controls for this so-called overfitting by a pruning procedure, which prunes the maximal tree back to some optimal subtree. The procedure relies on the fact that the QUINT algorithm yields a sequence of nested subtrees, which differ in number of leaves. For each of these subtrees, QUINT has computed a criterion value. Yet, these values are positively biased since they have been computed on the basis of the same data as the ones that were used to build the tree. To overcome this problem, QUINT relies on a bootstrap procedure that yields an estimation of the biases. Generally speaking, the (nonparametric) bootstrap is a procedure to calculate measures of accuracy (e.g., bias, confidence interval, prediction error) of model estimates by making use of a series of samples that are obtained by resampling with replacement from the original data (Efron, 2003). The bootstrap estimation of the biases further allows for the calculation of bias-corrected criterion values. The final (sub)tree selected by QUINT then is the one with the highest bias-corrected criterion value. QUINT also allows for a bootstrap-based bias correction procedure for the differential treatment effect sizes in the leaves of the finally selected tree, which may give insight into the generalizability of the QUINT solution. For a formalization and detailed description of the above, we refer to Dusseldorp and Van Mechelen (2014).

Results of a simulation study suggested that QUINT firstly has an overall good optimization performance in terms of maximizing the QUINT criterion function (Dusseldorp & Van Mechelen, 2014). A second group of performance criteria addressed in the simulation study pertains to recovery of the true structure underlying the data. These performance criteria include both correctness of inferences regarding the presence/absence of qualitative interactions (i.e., Type I and Type II error rates), and recovery of the structure of the underlying true tree. Simulation results of the Type I and Type II error indicated that for sample sizes larger than 300 “a good balance between the two can be obtained if a value of $d_{\min} = 0.30$ is used in the qualitative interaction condition” (Dusseldorp & Van Mechelen, 2014, p. 231). Regarding the recovery of aspects of the true underlying tree such as tree complexity, splitting variables and split points, and assignment of the clients to the three subgroups, the simulation study revealed that satisfactory results are obtained when the sample size is at least 400 and the true differences in treatment outcome in the subgroups are large (Cohen's $d \geq |1|$). To clarify, when no qualitative treatment–subgroup interaction is found in a situation where the sample size is 400 and $d_{\min}$ is set to 0.30, there is a reasonable basis for concluding that the moderators under study are not involved in a qualitative treatment–subgroup interaction; conversely, in case a qualitative treatment–subgroup interaction is found in such a situation, there is a reasonable basis for concluding that this interaction holds, indeed, and that the identified tree closely resembles the underlying true tree. Note that these recommendations are based on simulation results of Dusseldorp and Van Mechelen (2014) for data sets with a number of potential treatment moderators ranging from 5 to 20. In this simulation study the impact of the number of moderators on Type I and Type II error rates appeared to be negligible. That being said, we cannot exclude that in the case of a much larger number of potential moderators, the required sample size may be higher. Note finally also that the recovery results of the simulation study of Dusseldorp and Van Mechelen (2014) imply that QUINT analyses of different RCTs (in which the same treatment alternatives are compared while measuring the same outcome variable and the same background characteristics) should, in principle lead to similar results.

To analyze RCTs with QUINT, the R-package quint has been developed by Dusseldorp, Doove, and Van Mechelen (2013).[1] This package can be freely downloaded from CRAN. The current version of QUINT can handle two treatment alternatives only.

Table I. Percentage or mean (and standard deviation) for all potential moderators involved in reanalysis of data from the Clinical Trials Network. All potential moderators were measured before treatment.

| | Percentage or mean (*SD*) | |
|---|---|---|
| Potential moderator | Standard ($n = 214$) | MI ($n = 209$) |
| Female | 42.5 | 24.0 |
| Ethnicity | | |
| White | 70.6 | 73.2 |
| Other | 16.4 | 15.3 |
| Black, African American, or Negro | 9.3 | 9.6 |
| Spanish, Hispanic, or Latina | 3.7 | 1.9 |
| Employed | 40.2 | 34.0 |
| Marital status | | |
| Divorced | 18.7 | 25.4 |
| Living with partner/cohabiting | 2.8 | 2.9 |
| Separated | 10.7 | 9.6 |
| Legally married | 16.4 | 17.7 |
| Never married | 50.9 | 43.5 |
| Widowed | 0.5 | 1.0 |
| Admission prompted by legal system | 54.2 | 52.2 |
| On probation or parole | 39.3 | 35.9 |
| Any previous drug/alcohol treatment | 63.1 | 60.3 |
| Primary drug used | | |
| Alcohol | 47.7 | 48.8 |
| Cocaine | 7.0 | 5.7 |
| Marijuana | 21.0 | 21.1 |
| Opiates | 4.2 | 5.7 |
| Methamphetamines | 19.2 | 18.1 |
| Benzodiazepines | 0.9 | 0.5 |
| Age | 32.4 (9.7) | 34.3 (10.3) |
| Years of education | 12.1 (2.1) | 12.2 (1.7) |
| Days of substance use, past 30 | 10.1 (9.6) | 11.9 (10.6) |
| ASI composite scores | | |
| Medical | 0.27 (0.35) | 0.27 (0.35) |
| Employment | 0.67 (0.31) | 0.69 (0.31) |
| Alcohol | 0.23 (0.25) | 0.25 (0.27) |
| Drug | 0.11 (0.12) | 0.12 (0.12) |
| Legal | 0.20 (0.22) | 0.19 (0.21) |
| Family | 0.17 (0.21) | 0.16 (0.20) |
| Psychological | 0.25 (0.23) | 0.26 (0.24) |

*Notes:* Composite scores were calculated according to McGahan, Griffith, Parente, and McLellan (1986) and have a theoretical range from 0.00 to 1.00. *SD*, standard deviation; MI, motivational interviewing; and ASI, addiction severity index.

## Data

We re-analyzed data from the Clinical Trials Network[2] on the evaluation of integrating motivational interview techniques into the initial contact and evaluation session of behavior therapies (Carroll et al., 2006). Motivational interviewing has been developed as a treatment strategy to enhance clients' motivation for change (Miller & Rollnick, 1991). It includes both a motivational interviewing style (e.g., asking open-ended questions, listening reflectively, and affirming change-related participant statements and efforts), and motivation-enhancing strategies (e.g., practicing empathy, providing choice, clarifying goals). The data pertain to an RCT with participants ($n = 423$) who were seeking treatment for a substance use problem. Following baseline assessment, participants were randomly assigned to one out of two conditions: standard intervention ($n = 214$) and standard intervention in which motivational interviewing techniques were integrated in the intake/orientation sessions ($n = 209$). The data comprised 18 pre-treatment characteristics, such as

Table II. Available cases and mean (and standard deviation) for the outcome variables involved in reanalysis of data from the Clinical Trials Network.

| | Available cases | | Mean (*SD*) | |
|---|---|---|---|---|
| Outcome | Standard | MI | Standard | MI |
| Number of sessions in 28 days after treatment assignment | 178 | 174 | 4.1 (4.1) | 5.0 (5.1) |
| Number of days of substance use in 28 days after treatment assignment | 178 | 173 | 3.0 (6.2) | 3.4 (6.9) |

*Note*: *SD*, standard deviation and MI, motivational interviewing.

demographical variables (e.g., gender, age, ethnicity) and aspects of substance use (e.g., days of substance use in 30 days before treatment assignment, the primary drug used, and composite scores included in the Addiction Severity Index (ASI; McLellan et al., 1992), the latter being an interview-based measure of the frequency and severity of substance use and related psychosocial problems). Descriptive statistics for all pre-treatment characteristics are given in Table I. As outcome variables, we focused in our re-analyses on a measure of retention (a variable closely linked to motivation for change; Ryan, Plant, & O'Malley, 1995), and a measure of substance use. More specifically, the outcome variables were the number of therapy sessions completed and the number of days on which the participant reported using her or his identified primary substance of abuse, both during the 28 days after treatment assignment. The number of available cases for analysis (due to missing values at evaluation) and descriptive statistics for the two outcome variables are given in Table II. Two remarks can be made with regard to the missing values at evaluation, which should be kept in mind when analyzing the data. The first pertains to the number of available cases due to missing values, which falls below the required 400 to safeguard recovery of structural aspects of the underlying true tree. The second pertains to the missing data mechanism. That is, in the context of the data at hand, the reason why a value is missing is likely to be related to the value of the variable that is missing (technically speaking, missing not at random).

Previous analyses of these data (Carroll et al., 2006) showed that integrating motivational interviewing techniques in a standard treatment tends to have a positive effect on retention in the earlier phases of treatment ($d = 0.24$), but has no significant effect on substance use. Carroll et al., however, also hypothesized that treatment effect heterogeneity may be in place, and therefore that "it is important to understand the types of individuals for whom motivational interviewing is effective … " (p. 310). Conversely, a recent meta-analysis of 34 studies on motivational interviewing that used a measure of treatment engagement (e.g., keeping appointments, participation in treatment) as an outcome variable, showed that clients receiving motivational interviewing were not significantly advantaged over those who received a standard intervention (Lundahl, Kunz, Brownell, Tollefson, & Burke, 2010). Moreover, with regard to substance use-related outcomes, the meta-analysis also showed that clients receiving motivational interviewing were not significantly advantaged over those who received a standard intervention. However, the individual studies included in the meta-analysis showed a wide variability in the size and, more importantly, the direction of the differential treatment effects. This variability in outcomes across studies points to the possibility of treatment effect heterogeneity and, in particular, of qualitative treatment–subgroup interactions. This is exactly the issue that is addressed by QUINT in terms of identifying subgroups for which motivational interviewing outperforms standard treatment, whereas for the other subgroups the reverse holds true.

## Results of QUINT reanalyses and discussion

Given that the outcome variables (number of completed sessions, number of days of substance use) had a clear pragmatic meaning, the QUINT analyses were performed using the criterion with difference in treatment means. We set the number of bootstrap samples equal to 200 and used the default values for the weights of the two constituents of the criterion, the critical minimum value in the qualitative interaction criterion for the absolute value of the standardized mean difference in treatment outcome ($d_{\min} = 0.30$), the maximum number of leaves (10), and the minimum number of clients assigned to treatment A and B in each leaf (10% of the total number of clients assigned to treatment A and B).

### Retention

Regarding the number of completed sessions in the 28 days after treatment assignment, the test of the qualitative interaction condition revealed that a qualitative treatment-subgroup interaction is present in the data. QUINT subsequently constructed a tree with six leaves. The pruning procedure indicated that this was also the optimal tree size. After applying the bias correction procedure to the two leaves with most extreme differential treatment effects, we found on the one hand a leaf with clients who completed on average 2.60 more sessions during the 28 days after treatment assignment when assigned to motivational interviewing compared to standard treatment ($d = 0.42$), and, on the other hand, a leaf with clients who completed on average 2.84 more sessions when assigned to standard treatment compared to motivational interviewing ($d = -0.54$) (with the values of the effect sizes implying that the detected qualitative treatment–subgroup interaction is of medium size).

The structure of the full tree is shown in Figure 2. The ellipses in the figure represent the internal nodes containing the split variables, with the corresponding split point shown below each ellipse. The upper ellipsis represents the root node, which corresponds to the complete group of clients. The rectangles represent the leaves of the tree, that is, the final subgroups of clients; each rectangle contains the sample size of the corresponding subgroup, the outcome means (and standard deviations) for the motivational interviewing and the standard treatment condition ($\overline{Y}_{\text{MI}}$ and $\overline{Y}_{\text{standard}}$), the (uncorrected) difference in means ($\overline{Y}_{\text{MI}} - \overline{Y}_{\text{standard}}$), and the corresponding effect size $d$. For example, clients with an ASI composite score for drug use strictly larger than 0.26 end up in Leaf 6. Clients in this leaf completed on average 4.66 more therapy sessions after standard treatment compared with motivational interviewing; consequently, this leaf is assigned to subgroup $\wp_2$.

Looking at the QUINT result as a whole, it appears that users with more severe drug problems (drug use ASI composite score $> 0.26$) take more advantage from standard treatment compared to motivational interviewing. On the other hand, male users with less severe drug problems (drug use ASI composite score $\leq 0.26$), who used drugs during an intermediate number of days before treatment assignment (2 to 19 days out of 30) should preferably receive motivational interviewing (as they then complete on average 5.30 more sessions than when assigned to standard treatment). The latter also holds for women if their employment strengths are not too heavily affected (with then on average 3.41 more completed sessions after motivational interviewing).

From the above, we may conclude that the most important moderators of the differential effectiveness of motivational interviewing and standard treatment are measures of problem severity (i.e., drug use ASI composite score, days of substance use, lack of employment strengths) and gender. Regarding problem severity, this result somewhat links up with the meta-analysis by Lundahl et al. (2010), who found an (albeit nonsignificant) trend for clients' level of impairment to moderate the overall effect of motivational interviewing in the same direction as we did (i.e., less impaired clients profit more from motivational interviewing). Furthermore, in other areas, differential treatment outcomes for clients with high and low problem severity have been reported as well (e.g., Elkin et al., 1995). Regarding gender, as noted by Green (2006) in a discussion of substance abuse treatment services, "Researchers also have identified many factors that differ by gender and affect treatment outcomes in important ways ( … ). This suggests that addressing risks differentially, by gender, may help improve both the treatment process and outcomes for men and women" (p. 60). Finally, although one should be cautious with inferences about structural aspects of the tree as the sample size was smaller than 400, it is interesting to note that the split point of 0.26 on the ASI composite score for drug use that was identified by QUINT, is theoretically meaningful, as it almost coincides with the critical value of 0.25 that Lee et al. (2001) used to mark off a group of users with severe drug use problems.

One may finally note that the conclusion of Carroll et al. (2006) that, when assigned to motivational interviewing, clients have better retention through the 28 days after treatment assignment, is to be significantly qualified on the basis of the QUINT results: Motivational interviewing appears not to be the preferred choice of treatment for substance users with more severe drug problems. At first sight, the finding that both a main effect of motivational interviewing and a qualitative treatment–subgroup interaction are present in the data might look somewhat contradictory. However, this pattern can be readily explained by the fact that the group of users who take more advantage from motivational interviews is considerably larger than the group of users for whom standard treatment is the preferred treatment alternative.

## Substance Use

Regarding the outcome variable "Number of days of substance use during the 28 days after treatment assignment," the test of the qualitative interaction condition in the root node revealed that no qualitative treatment–subgroup interaction was present in the data (with $d = -0.25$ and $d = 0.28$ in the leaves). Carroll et al. (2006) have previously concluded that there is no significant positive effect of integrating motivational interviewing techniques into the initial contact and evaluation session of behavior therapies on frequency of substance use. Based on the QUINT result, no indications are found that this does not apply to all substance users. Moreover, the fact that in this case QUINT did not identify a qualitative treatment–subgroup interaction also nicely illustrates the inbuilt protection within the QUINT procedure to safeguard against erroneous claims about apparent interactions that cannot be replicated in follow-up studies.

To summarize, the combination of the QUINT results regarding the number of completed sessions and the number of days of substance use in the 28 days after treatment assignment leads to a personalized treatment assignment strategy that implies a gain in the number of completed sessions (i.e., increased retention), yet without a sizeable decrease in substance use. This may be considered a modest return.

## Concluding Remarks

In this paper we reviewed a recently developed, powerful tool for the identification of subgroups of clients that are involved in clinically meaningful qualitative treatment–subgroup interactions. This tool has been devised for the most common context of RCTs with large numbers of possible moderators and without comprehensive a priori hypotheses on the subgroups that are subject to differences in differential treatment effectiveness. An important advantage of QUINT is that it may lead to insightful and well-interpretable results, despite the fact that the treatment–subgroup interactions at hand may rely on a complex interplay of moderators, with individual moderators possibly even being involved in the interactions in a nonlinear way (as was the case in our application with the variable “Number of days with substance use before treatment assignment”). Secondly, and even more importantly, the QUINT results may have straightforward implications for personalized treatment assignment.

An important problem in the study of treatment–subgroup interactions is the potential for sizeable inferential errors (Rothwell, 2005). As such, QUINT incorporates several tools to control for inferential errors. Firstly, the risk of erroneously concluding that a qualitative interaction is present in the data is met by the test of the qualitative interaction criterion in the root node. Secondly, overfitting (and, hence, arriving at conclusions that hold for the data at hand only and not for future data) is controlled for via the model selection (i.e., stop and pruning criteria) and through the calculation of bias-corrected values of the effects and the effect sizes in the leaves. Thirdly, an extensive simulation study led to several guidelines to assess the risk of inferential errors (Dusseldorp & Van Mechelen, 2014). As an example of the latter, our application of the guidelines suggested that we could safely conclude that a qualitative treatment–subgroup interaction is present in the data on retention in the earlier phases of treatment; yet, taking into account a sample size below 400 and the fact that the bias correction procedures suggested that the detected interaction was moderate in size, they also imply that we should be cautious with regard to inferences about the complexity and the structure of the underlying true tree. The latter implies an important additional incentive for something that should preferably be done after exploratory subgroup analyses such as QUINT, that is, examine whether the found treatment-subgroup interactions can be replicated. Ideally the output of QUINT should be used for setting up new trials using stratified randomizations, with strata that are constructed on the basis of the subgroups identified by QUINT.

A broad concern related to the goal of finding subgroups involved in treatment-subgroup interactions is that, in general, a reliable detection of interactions requires larger samples than a reliable detection of main effects (Lee et al., 2015), and perhaps considerably larger than those enrolled in traditional clinical trials in the field of psychotherapy research. This, however, does not preclude that in such trials the use of subgroup analyses such as QUINT can be meaningful. As noted earlier, conducting an RCT requires lots of time, money and effort, which is mainly an argument for getting as much information as possible out of the data. Exploratory analyses on smaller data sets can, when reported as such, still be of great value. However, it is of utmost importance that the results of them are regarded as tentative until they can be replicated.

Admittedly, the QUINT method has a number of limitations by itself. First, the criterion that is optimized by QUINT is an ad hoc criterion that is not at the level of inferences about a population. Second, QUINT, as most tree-based methods, is estimated on the basis of a greedy heuristic. Third, QUINT is limited to the case of two alternative treatments, whereas quite a few RCTs include more than two arms. The development of a new method for the detection of qualitative treatment–subgroup interactions that overcomes these three limitations looks like an important challenge for future research.

## Funding

The research reported in this paper was supported in part by the Fund for Scientific Research Flanders

(G.0546.09N), by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles program financed by the Belgian government (IAP/P7/06). The paper reported results from secondary analyses of data from clinical trials conducted as part of the National Drug Abuse Treatment Clinical Trials Network (CTN) sponsored by National Institute on Drug Abuse (NIDA). Specifically, we reanalyzed data from CTN-0005 entitled "MI (Motivational Interviewing) to Improve Treatment Engagement and Outcome in Subjects Seeking Treatment for Substance Abuse."

## Notes

[1] The package posted on Comprehensive R Archive Network (CRAN) can handle continuous and dichotomous background characteristics. For the analyses in the present paper, we used a slightly extended version of this package that can also handle categorical background characteristics. This extension can be obtained from the first author.

[2] Clinical Trials Network databases and information are available at www.ctndatashare.org.

## References

Bala, M. M., Akl, E. A., Sun, X., Bassler, D., Mertz, D., Mejza, F., … Guyatt, G. H. (2013). Randomized trials published in higher vs. lower impact journals differ in design, conduct, and analysis. *Journal of Clinical Epidemiology*, *66*, 286–295. doi:10.1016/j.jclinepi.2012.10.005

Boonacker, C., Hoes, A., Van Liere-Visser, K., Schilder, A., & Rovers, M. (2011). A comparison of subgroup analyses in grant applications and publications. *American Journal of Epidemiology*, *174*, 219–225. doi:10.1093/aje/kwr075

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Byar, D. P. (1985). Assessing apparent treatment—covariate interaction in randomized clinical trials. *Statistics in Medicine*, *4*, 255–263.

Carroll, K. M., Ball, S. A., Nich, C., Martino, S., Frankforter, T. L., Farentinos, C., … National Institute on Drug Abuse Clinical Trials Network (2006). Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: A multisite effectiveness study. *Drug and Alcohol Dependence*, *81*, 301–312. doi:10.1016/j.drugalcdep.2005.08.002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684. doi:10.1037/h0043943

De Jong, K., Timman, R., Hakkaart-Van Roijen, L., Vermeulen, P., Kooiman, K., Passchier, J., & Busschbach, J. V. (2014). The effect of outcome monitoring feedback to clinicians and patients in short and long-term psychotherapy: A randomized controlled trial. *Psychotherapy Research*, *24*, 629–639.

Dixon, D. O., & Simon, R. (1991). Bayesian subset analysis. *Biometrics*, *47*, 871–881. doi:10.2307/2532645

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find subgroups involved in meaningful treatment-subgroup interactions. *Advances in Data Analysis and Classification*, *8*, 403–425. doi:10.1007/s11634–013-0159-x

Dusseldorp, E., Doove, L. L., & Van Mechelen, I. (2013). Quint: Qualitative interaction trees. R Package Version 1.0. Retrieved from http://cran.r-project.org/package=quint

Dusseldorp, E., & Van Mechelen, I. (2014). Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, *33*, 219–237. doi:10.1002/sim.5933

Efron, B. (2003). Second thoughts on the bootstrap. *Statistical Science*, *18*, 135–140. doi:10.1214/ss/1063994968

Elkin, I., Gibbons, R. D., Shea, M. T., Sotsky, S. M., Watkins, J. T., Pilkonis, P. A., & Hedeker, D. (1995). Initial severity and differential treatment outcome in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology*, *63*, 841–847. doi:0022-006X/95/J3.00

Feinstein, A. R. (1998). The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology*, *51*, 297–299. doi:10.1016/S0895-4356(98)00004-3

Green, C. A. (2006). Gender and use of substance abuse treatment services. *Alcohol Research and Health*, *29*, 55–62.

Gullestad, F. S., Johansen, M. S., Høglend, P., Karterud, S., & Wilberg, T. (2013). Mentalization as a moderator of treatment effects: findings from a randomized clinical trial for personality disorders. *Psychotherapy Research*, *23*, 674–89. doi:10.1080/10503307.2012.684103

Hayward, R. A., Kent, D. M., Vijan, S., & Hofer, T. P. (2006). Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology*, *6*, 18. doi:10.1186/1471-2288-6-18

Kent, D. M., Rothwell, P. M., Ioannidis, J. P. A., Altman, D. G., & Hayward, R. A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials*, *11*, 85. doi:10.1186/1745-6215-11-85

Lee, P. A., Marlowe, D. B., Festinger, D. S., Cacciola, J. S., McNellis, J., Schepise, N. M., & McLellan, A. T. (2001). Did "Breaking the Cycle" (BTC) clients receive appropriate services? [abstract]. *Drug and Alcohol Dependence*, *63*(Suppl. 1), S89 [Presentation at the 63rd Annual Scientific Meeting of the College on Problems of Drug Dependence; Scottsdale, AZ].

Lee, S., Lei, M.-K., & Brody, G. H. (2015). Confidence intervals for distinguishing ordinal and disordinal interactions in multiple regression. *Psychological Methods*, 20, 245–258. doi:10.1037/met0000033

Lundahl, B. W., Kunz, C., Brownell, C., Tollefson, D., & Burke, B. L. (2010). A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on Social Work Practice*, *20*, 137–160. doi:10.1177/1049731509347850

McGahan, P. L., Griffith, J. A., Parente, R., & McLellan, A. T. (1986). *Addiction severity index composite scores manual*. Philadelphia, PA: Treatment Research Institute.

McLellan, A. T., Kushner, H., Metzger, D., Peters, R., Smith, I., Grissom, G., … Argerious, M. (1992). The fifth edition of the addiction severity index. *Journal of Substance Abuse Treatment*, *9*, 199–213. doi:10.1016/0740-5472(92)90062-S

Miller, W. R., & Rollnick, S. (1991). *Motivational interviewing: Preparing people to change addictive behavior*. New York, NY: Guilford Press.

Ogrodniczuk, J. S., Piper, W. E., Joyce, A. S., & McCallum, M. (2001). Effect of patient gender on outcome in two forms of short-term individual psychotherapy. *Journal of Psychotherapy Practice and Research*, *10*, 69–78.

Peto, R. (1995). Clinical trials. In P. Price & K. Sikoa (Eds.), *Treatment of cancer* (pp. 1039–1043). London: Chapman & Hall.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*, *21*, 2917–2930. doi:10.1002/sim.1296

Rothwell, P. M. (2005). Subgroup analysis in randomized controlled trials: Importance, indications, and interpretation. *The Lancet*, *365*, 176–186. doi:10.1016/S0140-6736(05)17709-5

Ryan, R. M., Plant, R. W., & O'Malley, S. (1995). Initial motivations for alcohol treatment: relations with patient characteristics, treatment involvement, and dropout. *Addictive Behaviors*, *20*, 279–297.

Shaffer, J. P. (1991). Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika*, *56*, 29–38. doi:10.1007/BF02294583

Tasca, G. A., Ritchie, K., Conrad, G., Balfour, L., Gayton, J., Lybanon, V., & Bissada, H. (2006). Attachment scales predict outcome in a randomized controlled trial of two group therapies for binge eating disorder: An aptitude by treatment interaction. *Psychotherapy Research*, *16*, 106–121. doi:10.1080/10503300500090928

Tunis, S. R., Benner, J., & McClellan, M. (2010). Comparative effectiveness research: Policy context, methods development and research infrastructure. *Statistics in Medicine*, *29*, 1963–1976. doi:10.1002/sim.3818

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine: Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, *357*, 2189–2194. doi:10.1056/NEJMsr077003