Published in final edited form as:

Am J Drug Alcohol Abuse. 2015 ; 41(6): 498–507. doi:10.3109/00952990.2015.1044605.

A Systematic Approach to Subgroup Analyses in a Smoking Cessation Trial

A. N. Westover^{1,2}, T. M. Kashner^{1,3,4}, T. M. Winhusen⁵, R. M. Golden⁶, PA Nakonezny^{1,2}, B. Adinoff^{1,7}, and S.S. Henley^{3,8}

¹Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX

²Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX

³Department of Medicine, Loma Linda University School of Medicine, Loma Linda, CA

⁴Veterans Health Administration Office of Academic Affiliations, Washington, D.C.

⁵Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati College of Medicine, Cincinnati, OH

⁶University of Texas at Dallas, School of Behavioral and Brain Sciences, Richardson, TX

⁷VA North Texas Health Care System, Dallas VAMC, Dallas, TX

⁸Martingale Research Corporation, Plano, TX

Abstract

Background—Traditional approaches to subgroup analyses that test each moderating factor as a separate hypothesis can lead to erroneous conclusions due to the problems of multiple comparisons, model misspecification, and multicollinearity.

Objective—Demonstrate a novel, systematic approach to subgroup analyses that avoids these pitfalls.

Methods—A Best Approximating Model (BAM) approach that identifies multiple moderators and estimates their simultaneous impact on treatment effect sizes was applied to a randomized, controlled, 11-week, double-blind efficacy trial on smoking cessation of adult smokers with ADHD, randomized to either OROS-methylphenidate (n=127) or placebo (n=128), and treated with nicotine patch. Binary outcomes measures were prolonged smoking abstinence and point prevalence smoking abstinence.

Results—Although the original clinical trial data analysis showed no treatment effect on smoking cessation, the BAM analysis showed significant subgroup effects for the primary outcome of prolonged smoking abstinence: 1) lifetime history of substance use disorders (adjusted odds ratio [AOR] 0.27; 95% confidence interval [CI] 0.10–0.74), and 2) more severe ADHD symptoms (baseline score >36; AOR 2.64; 95% CI 1.17–5.96). A significant subgroup effect was

Address correspondence to Arthur Westover, Departments of Psychiatry and Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390. arthur.westover@utsouthwestern.edu.

Declaration of Interests

The authors report no conflicts of interest.

also shown for the secondary outcome of point prevalence smoking abstinence—age 18 to 29 years (AOR 0.23; 95% CI 0.07–0.76).

Conclusions—The BAM analysis resulted in different conclusions about subgroup effects compared to a hypothesis-driven approach. By examining moderator independence and avoiding multiple testing, BAMs have the potential to better identify and explain how treatment effects vary across subgroups in heterogeneous patient populations, thus providing better guidance to more effectively match individual patients with specific treatments.

Keywords

subgroup analysis; statistics; modeling; tobacco; methylphenidate; Attention Deficit Hyperactivity Disorder

BACKGROUND

Clinical scientists often apply subgroup analyses to randomized clinical trials (RCT) data to better understand how treatment effects vary by patients (1). Such analyses enable clinicians to provide patient-centered, personalized medicine by matching correct treatments with individual patients (2). Preferably, only patients whose benefits exceed risk of side effects should receive treatment (3). In this way, subgroup analyses may guide treatment recommendations for specific patient subgroups (4).

Studies applying subgroup analyses often test each subgroup as a separate hypothesis. These hypothesis-driven approaches present problems in the absence of a widely accepted clinical theory. On one hand, considering many different groups can lead to spurious results due to inflated type I error from multiple comparisons. On the other hand, focusing on a limited number of groups may prevent discovering novel associations that could provide new findings.

Providing criteria to conduct hypothesis-driven subgroup analyses in clinical trials, Sun et al (5). advocate (a) testing no more than five prespecified subgroups using background variables that are stratified at randomization, (b) prespecifying the direction of the effect, (c) testing a subgroup effect as an interaction term while controlling for other interactions with other moderators, (d) determining consistency of subgroup effects across related outcomes, and finally (e) performing a qualitative evaluation of the subgroup effect. Unfortunately these prescribed criteria are rarely applied, and remain subject to publication bias that occurs when only studies with statistically significant findings are reported in the literature.

As an alternative to hypothesis-driven subgroup analyses, Wallace et al. collapsed 32 potential baseline moderators into a single index that could be tested for its impact on treatment effect sizes (6). Specifically, the number of moderators was reduced to eight using “individual moderator effect sizes, loadings from the principal-components analysis, clinical meaningfulness, and access to complete data.” Weights were calculated for each of these eight moderators to form a combined moderator score, or M^* that has been shown to discriminate outcomes between treatment options. However, the authors identified two problems: (a) the combined moderator is only one possible model among reasonable alternatives, and (b) interactions between individual moderators were not considered.

Building on the Wallace et al. single index moderator score, we describe a specification-driven approach that focuses on finding and testing specific moderators on treatment effect sizes. Rather than constructing moderator scores as an intermediate model, our premise is that we assess effect size and moderator scores simultaneously within a single outcome model.

To identify and estimate the “true” data generating process (DGP) of patient outcomes, we use a principled, systematic, and transparent strategy to estimate a Best Approximating Model, or BAM, that is selected from among all possible models to have the best “fit” to study data using model selection criteria (7–9). This strategy enables investigators to consider simultaneously more expansive sets of potential moderators and covariates (25). Further, BAM is designed to improve upon hypotheses-driven approaches that can consider only a limited number of moderators and covariates (5), and a moderator-score approach that severely limits the search space of possible models.

BAM is very different than other model selection procedures with which most investigators are familiar, such as stepwise regression. In contrast to the BAM method, stepwise selection approaches consider only a small percentage of candidate models from the set of all possible models. In addition, stepwise procedures as commonly implemented in existing statistical software use a biased sequential model selection process, do not control for effects of model misspecification, and employ hypothesis testing methods such as the likelihood ratio test that do not control for experiment-wise error rates. In the BAM method, a large percentage of the model space is systematically explored and the model search procedure is not biased. In addition, unbiased model selection criteria such as GAIC (9) are used rather than hypothesis testing procedures for model comparison. As a result of these strategies used in the BAM method, a model space is explicitly defined and then all models in that model space are ranked to find the best model using GAIC, which controls for both model complexity (i.e., overfitting) and model misspecification.

We applied BAM to a National Institute on Drug Abuse (NIDA) Clinical Trials Network (CTN) treatment study of smoking cessation. The primary hypothesis of this randomized double-blind placebo-controlled intent-to-treat study was that adults with Attention Deficit Hyperactivity Disorder (ADHD) would have improved smoking cessation if treated for ADHD symptoms with the stimulant medication osmotic-release oral system methylphenidate (OROS-MPH). In the original study (10), there was no effect of treatment of OROS-MPH on smoking cessation outcomes despite prior research suggesting that such effects could be present (10;11). This study aimed to provide additional insights into understanding factors influencing smoking cessation outcomes by performing the BAM analysis and comparing those results with a contemporary hypothesis-driven approach.

METHODS

Participants and Procedures

As described elsewhere (12), 255 currently-smoking adults ranging from 18–55 years of age who also met DSM-IV criteria for ADHD were randomized to either OROS-MPH (n=127) or placebo (n=128). All participants received transdermal nicotine patches and weekly

smoking cessation counseling. The 11-week trial consisted of a pre-quit phase, designated quit-smoking day, and post quit phase. Since smoking is more prevalent among adults with ADHD than the general population (13;14) and may reflect self-treatment of ADHD symptoms (15), it was reasoned that the pharmacologic treatment of ADHD in adults with comorbid ADHD and nicotine dependence may attenuate smoking. Study completion (84%) and medication compliance (94% pills taken) were high. The multi-site study (2005–2008) was registered with <http://clinicaltrials.gov> (NCT00253747) and was approved by institutional review boards at participating sites.

Measures

Outcome Variables—As in the CTN trial, the primary outcome variable was prolonged smoking abstinence (16) i.e. self-report of tobacco abstinence without treatment failure (defined as smoking each day for 7 consecutive days or having smoked at least 1 day of each week in 2 consecutive weeks) during study weeks 7–10. This binary variable assumes the value of one (i.e. treatment success) and zero (i.e. treatment failure) otherwise. The secondary outcome, point prevalence abstinence assumes the value of one (i.e. treatment success) if during study week 10 the patient’s self-report of not smoking during a prior seven-day period was confirmed by a CO level <8 ppm, and zero (i.e. treatment failure) otherwise.

Treatment Indicator—The binary treatment indicator variable assumes the value of one if the patient was randomized to OROS-MPH, and zero if randomized to placebo.

Study Covariates—Potential covariates of smoking cessation were identified a priori (Table 1). The demographic and health variables selected as covariates reflected a broad inclusion of plausible variables in order to not exclude important and/or novel models. Demographic covariates included age, gender, race/ethnicity (non-Hispanic White vs. not), marital status, years of education, and employment status. Baseline health covariates included body weight, severity of nicotine dependence based on the Fagerstrom test score (17) (higher scores reflecting greater dependence), number of prior smoking quit attempts, start age for smoking, lifetime history of psychiatric disorders, baseline ADHD Rating Scale-IV total score with higher scores indicating greater illness severity (18), baseline Beck Depression Inventory (BDI)-II (19) and Beck Anxiety Inventory (BAI) (20) scores (higher scores indicating more severe symptoms), and ADHD type (“hyperactive/impulsive” type was added to the “combined” type due to the small number of the former). The study site covariate (six sites) was recoded using a binary design variable. For continuous/ordinal variables that were not linear in the logit as determined by Generalized Information Matrix Tests (21) and the Box-Tidwell link specification test (22), cutpoints were clinically prespecified for binary recoding. Accordingly, cutpoints for BDI (14) and BAI (8) were chosen to indicate at least mild depression and mild anxiety, respectively. Years of education was dichotomized at 13, to signify presence or absence of more than a high school education. ADHD Rating Scale-IV was dichotomized at > 36 per Nunes et. al (2013) (23). Age was trichotomized at standard intervals (the oldest participant was 56 years old). A median value was prespecified as the cutpoint for start age for smoking and number of prior smoking quit attempts.

Analyses

Best Approximating Model—BAM subgroup analysis is designed to find the clinical trial data generating process (DGP) that contains the true relationships between the patient's health outcomes, and the treatment indicator, covariates (predictors), and treatment indicator-covariate interactions (moderators). If present, coefficients to interaction terms measure the impact of the corresponding moderator covariate on treatment effect size. The coefficient to the treatment indicator term measures the treatment effect size when all moderator covariates are set at zero.

For purposes of assessing moderators to treatment effect sizes, a BAM is selected and estimated from a model space of candidate models that contain the treatment indicator variable, plus all possible combinations of covariates and indicator-covariate interactions. The following criteria were used to identify a BAM. First, the model includes the treatment indicator variable. Second, the correlation among its covariates and indicator-covariate interactions is low with condition numbers less than 1,000, where the condition number equals the square root of the maximum eigenvalue divided by the minimum eigenvalue of the estimated covariance matrix (collinearity). Third, there is no evidence the model is not correctly specified (24), where misspecification is tested using the Log Eigenspectrum and Log GAIC GIMTs (21) and the Box-Tidwell Link test (22) ($\alpha=0.05$). Fourth, the model is among the candidate approximating models defined as having the best ten-fold cross-validated fit based on Generalized Akaike Information Criterion (GAIC) (9). Akaike Information Criteria such as GAIC are designed to mathematically correct for overfitting effects without using cross-validation methods (8;9;25). Ten-fold cross-validation (26) using GAIC is utilized to provide further protection against over-fitting effects in the final model. And finally, the BAM is selected from among the candidate approximating models with the best ten-fold cross-validated estimate of GAIC.

Empirically identifying and estimating a BAM for each outcome variable requires three stages, as outlined in Figure 1.

Stage I (Covariate Pool): The purpose of the first stage is to create a pool of candidate covariates (predictors). We begin by selecting variables from the RCT dataset that will plausibly predict patient outcomes based on a review of the existing scientific literature, clinical experience, and theory. Each plausible variable may be transformed to best represent the DGP or facilitate interpretability. Specifically, categorical variables are transformed into a set of binary design variables (27). Continuous variables (e.g. ADHD Rating Scale-IV) that are not linear in the logit (21), are assigned prespecified cutpoints based on clinical criteria (28). Each of these variables and treatment interactions is checked separately in a univariate logistic regression model on the RCT dataset to identify variables that are not individually predictive of patient outcomes. Most variables have at least some predictive component, however variables identified as exhibiting evidence of no predictive performance were eliminated by using a likelihood ratio test (LRT) (29) with a significance level that was almost equal to one ($\alpha=0.99$). Finally, plausible, transformed, and predictive variables are entered into the covariate pool only if the univariate logistic regression model containing the candidate variable fails to test positive for misspecification.

Stage II (Candidate Approximating Models): In the second stage, we construct a set of treatment outcomes models that include the treatment indicator variable, plus all possible combinations of the final set of plausible, transformed, predictive, and specification-tested variables that comprise the covariate pool. We initially reduce the large covariate pool by conducting a stochastic search of randomly selected models and computing model fit using the Akaike Information Criterion corrected for small sample sizes (AICc) (8). Covariates and indicator-covariate interactions not represented in the top group models are eliminated. Next, all possible models are reconstructed from the remaining covariates in the reduced covariate pool. The set of candidate approximating models that have the best fit of the data based on GAIC is created by exhaustively searching (30–34) through these constructed models. Models having multicollinear predictors with condition numbers exceeding 1,000 are eliminated. A set of candidate approximating models is constructed from the remaining models by selecting the top 2,048 with the best cross-validated GAICs (8;9). It is important to emphasize that all 2,048 models at this stage of the process are considered to have essentially equivalent “model fit” to the underlying process that generated the observed data. Thus, additional criteria, other than model fit, must be used to identify a particular BAM.

Stage III (Best Approximating Model): In the third stage, the BAM is selected from among the pool of cross-validated candidate approximating models with the best ten-fold cross-validated GAIC. The best model is tested for misspecification using the Log Eigenspectrum and Log GAIC GIMTs ($\alpha=0.05$). If the model tests positive for misspecification, it is excluded and the model with the next best cross-validated GAIC performance is subjected to specification testing. The process continues until the best fitting model that does not test positive for misspecification is found. The BAM analyses were conducted using MATLAB 7.1 (MathWorks, Natick, MA). Standard errors on the final BAM are computed using robust covariance estimator (24).

Hypothesis-driven subgroup analyses—We also computed subgroup analyses for all covariates in the Stage I covariate pool. Models included the treatment indicator, the selected covariate, and the treatment-covariate interaction term whose coefficient estimates moderate effect size. Analyses were conducted using SAS software version 9.3 (SAS Institute, Inc., Cary, NC).

RESULTS

Demographics and Baseline Measures

Age, lifetime history of psychiatric disorders, and baseline measures for nicotine addiction (Fagerstrom test score), ADHD severity (ADHD Rating Scale total score), depression (BDI Score), and anxiety (BAI score) were similar between treatment groups (Table 1).

Primary Outcome—Prolonged Smoking Abstinence

The pool of recoded covariates represented a total of 1.126×10^{15} possible candidate logistic regression models. A stochastic search of 91,344,700 models reduced the number of candidate variables, leaving 524,288 models that were exhaustively searched. A 10-fold cross validation based on GAIC scores were computed for the top 2,048 exhaustively

searched models that had been ranked also by GAIC. A BAM was then selected as the model with the best cross-validated GAIC score. The AUROC for this model was 0.73 (95% confidence interval [CI], 0.67–0.80; Figure 2A).

Examining significant subgroup/OROS-MPH interactions for the BAM model, we found that more severe ADHD symptoms (ADHD-RS > 36) at baseline was associated with a 2.6-times improved likelihood of prolonged smoking abstinence, while conversely a lifetime history of a substance use disorder was associated with a 73% decreased likelihood of prolonged smoking abstinence (Table 2).

Secondary Outcome—Week 10 Point Prevalence Abstinence

The pool of recoded covariates represented a total 1.126E+15 possible candidate logistic regression models. A stochastic search of 99,466,000 candidate models reduced the number of candidate variables, leaving 524,288 models that were exhaustively searched. A 10-fold cross validation on GAIC scores was computed for the top 2,048 exhaustively searched models that had been ranked also by GAIC (9). A BAM was selected as the model with the best cross-validated GAIC score. The AUROC for this model was 0.73 (95% confidence interval [CI], 0.67–0.80; Figure 2B).

For this secondary outcome, the only significant subgroup/OROS-MPH interaction was for age 18 to 29 years, who had a 77% decreased likelihood of point prevalence smoking abstinence at week 10 (Table 3). A trend towards significantly improved abstinence ($p < 0.10$) was observed for participants with more severe ADHD symptoms. A clinical study site also evidenced significance as an adjusting covariate.

Hypothesis-Driven Subgroup Analyses

The hypothesis-driven subgroup analyses based on usual contemporary approaches demonstrated significant single-moderator subgroup associations ($p < .05$) between OROS-MPH treatment and prolonged smoking abstinence for ADHD symptom severity and ADHD subtype (Table 4). A trend towards significance ($p < 0.10$) was observed for lifetime substance use disorder. The only subgroup to show a signal with the secondary outcome (point prevalence abstinence) was ADHD severity ($p = 0.0147$).

DISCUSSION

BAM Results

This secondary analysis of CTN trial data, applied a specification-driven subgroup analyses that exhaustively searched for a Best Approximating “Outcome” Model (BAM) to investigate how OROS-MPH treatment impacts smoking abstinence when treatment effects are significantly influenced by moderating patient and clinical factors. While the original trial (10) showed no treatment benefit compared to placebo, BAM demonstrated that OROS-MPH effect sizes on smoking abstinence do exist, though vary across patient subgroups. Participants with no history of a substance use disorder or more severe ADHD symptoms had improved prolonged smoking abstinence with OROS-MPH treatment. For the secondary

outcome of point prevalence smoking abstinence, participants who were 18 to 29 years of age had a worsened outcome with OROS-MPH treatment.

BAM Versus Hypothesis Driven Subgroup Analysis

Unlike hypothesis-driven approaches, BAM offers a more comprehensive, systematic, and transparent approach to subgroup analyses. BAM enables all of the researcher specified candidate variables to be explicitly included in the covariate pool, a priori. BAM systematically accounts for independence between covariates without inflating type I error due to multiple comparisons.

In contrast, the hypothesis-driven approach often lacks transparency in the process of how some candidate hypotheses are selected for study while others are excluded from study. While some investigators use procedures such as stepwise regression, such methods consider only a severely limited subset of the entire, unconstrained model space. BAM on the other hand, selects the model that best represents the data generating process (DGP) in a manner analogous to the way statistical programs use maximum likelihood estimation to select the parameter estimates that identify the best model-fit to the data DGP for a particular model. With BAM, instead of searching for a parameter estimate in a space of possible parameter values, the stochastic/exhaustive search is for a best approximating model in a space of possible probability models that approximates the true DGP. In particular, BAM avoids multiple comparisons and enables investigators to measure simultaneously treatment effect sizes and the influence moderators play on effect size. Therefore, BAM addresses the need for improved methods that better evaluate the heterogeneity of treatments effects in clinical trials (35), or specifically "...whether we can identify the subgroups of patients who will benefit the most from (or are the most likely to be harmed by) specific interventions" (36).

For the primary outcome of prolonged smoking abstinence, BAM and hypothesis-driven findings were similar for ADHD severity ($p=0.020$ vs. $p=0.002$, respectively) and lifetime substance use disorder subgroups ($p=0.011$ vs. $p=0.098$). Results differed in that the hypothesis-driven approach found ADHD subtype to be a significant subgroup, while BAM did not. For the secondary outcome of point prevalence abstinence, BAM identified age as the only significant subgroup with trends for the subgroups of ADHD severity and study site, while the hypothesis-driven approach suggested only ADHD severity as an important subgroup.

BAM Versus Published Subgroup Analyses

Of the 18 subgroups considered in the BAM analysis, trial investigators planned analyses for 13 subgroups (Table 4). Analyses have been reported in four publications for 5 of those 13 subgroups. The authors found racial/ethnic differences, but used a novel outcome definition (37); a three-way interaction between ADHD subtype, nicotine dependence level (Fagerstrom test) and OROS-MPH treatment on prolonged smoking abstinence (38); and that large differences in prolonged abstinence by site-type were observed, but a site-type by treatment interaction was not (39). One study demonstrated that ADHD severity was not associated with a novel definition of abstinence (40). Another study showed the contrary--that ADHD severity was associated with prolonged abstinence--and was the one area of

overlapping agreement with BAM (23). Otherwise, differences in the findings between BAM and these published subgroup analyses was quite high. The previously published findings suggest race/ethnicity, nicotine dependence level, and ADHD subtype have subgroup effects. In contrast, BAM showed that race, nicotine dependence and ADHD subtype did not enter the best model. Moreover, BAM pointed to lifetime substance use disorder as an important moderator, whereas no published study did. Reasons for these differences include the use of different outcome definitions (37;40) and different covariates. As an example, one of the studies used a three-way treatment interaction to determine that ADHD subtype and nicotine dependence showed treatment differences (we limited BAM to two-way interactions between moderators and the treatment factor for this study) (38). None of these studies tested for independence of their findings with other subgroups. These examples highlight differences that are possible between a global and systematic approach (BAM) from hypothesis-driven subgroup analyses. Surely the case of this particular clinical trial is not unique.

Among our BAM findings, a history of substance use disorder predicting ineffective smoking cessation is well supported in the literature (41). While increased ADHD severity predicts an improved clinical response to OROS-MPH (42), no separate clinical trial has suggested that ADHD severity may be a moderator of OROS-MPH's effects on smoking cessation.

We suggest that our findings can be evaluated in the context of criteria proposed by Sun et al (5). Supporting BAM findings: (a) only baseline characteristics were used, (b) all moderators were specified a priori, (c) multiple testing was avoided, (d) tests for significant treatment interaction were used (43), (e) moderators were tested for independence, and (f) prior evidence was compatible with substance use disorder and ADHD severity moderator effects. Not supporting BAM findings: (a) analyzed moderators were not used as stratification factors at randomization, (b) the direction of subgroup effect was not prespecified (although the BAM method has this capability), and (c) the primary and secondary smoking cessation outcome measures did not show the same subgroup effects. Summarizing our understanding of the clinical problem, the clinical trial, and the BAM analysis, and having arguably achieved 6 of the 10 Sun et al. (5) criteria we find that a history of substance use disorder and more severe ADHD symptoms may be significant and important factors in OROS-MPH-mediated smoking cessation. However the relatively small sample size of the original clinical trial tempers this conclusion. Replication/confirmatory studies would be required to make a definitive judgment about the treatment effects of OROS-MPH in these subgroups.

Credibility of Subgroup Effect—Comparison of BAM and Sun et al.'s Criteria

In comparing BAM to Sun et al.'s criteria (5) (Table 5), there are a few important insights. First, in the design phase, BAM requires that investigators identify candidate covariates, instead of candidate hypotheses. In this way, BAM includes prior knowledge. Because BAM does not use multiple testing, but rather an information theoretic approach to model selection (8;9), the limitation on the number of subgroups to be analyzed is not required. This is an important advantage, especially when prior knowledge does not give adequate guidance in

the appropriate selection of subgroups. Second, in the analysis phase, BAM by design uses treatment-covariate interactions and tests for independence among all candidate subgroups. Contemporary hypothesis-driven approaches rarely examine independence among all subgroups (5). Third, in the context phase, BAM has no particular advantage or disadvantage compared to the hypothesis-driven approach—the results must be examined for robustness and compared to external evidence. Unlike contemporary approaches, BAM does not require the direction of treatment effect to be specified. After selection of the candidate subgroups, BAM imposes no other assumptions.

Conclusion

In this secondary study of a smoking cessation trial, the BAM approach identified subgroups of patients evidencing significant effects from OROS-MPH treatment. In contrast, the parent study (10) did not find an overall OROS-MPH treatment effect on smoking cessation. These divergent findings underscore the need for investigators to consider more advanced statistical methods to better analyze subgroup effect sizes.

Our findings also demonstrated the utility of a novel specification-based BAM approach for subgroup analyses of clinical trials data. The BAM approach offers improvements over traditional hypothesis-driven approaches by enabling investigators to systematically identify patient subgroups with associated treatment effects. In particular, the BAM approach: i) controls for type I error inflation by finding a single model using an information theoretic approach, ii) allows for consideration of a much larger group of possible covariates and moderator factors, iii) offers a systematic, principled, and replicable approach to identify and testing moderators, iv) estimates moderator effects after adjusting for the impact of all other covariates and moderators reported in the analysis, v) employs principled search algorithms that control for model specification and multicollinearity, and vi) utilizes robust statistics to estimate both model fit and test for model specification. This new approach allows investigators to determine how treatment effects may vary by patient.

In summary, while clinical investigators today may initially question the need for rigorously addressing model specification (44) when performing RCT analyses, we anticipate that future investigators will wonder how effect sizes were computed with statistical models where the possibility of misspecification was ignored. The BAM approach—made possible because of advances in computing and statistics—will be considered an essential robust subgroup analysis method for modeling patient-treatment heterogeneity in clinical trials.

Acknowledgments

Funding

This research was made possible by grants from the National Institute on Drug Abuse Clinical Trials Network (U10-DA013732, PI: E. Somoza); National Institute of General Medical Sciences (NIGMS) (R43GM106465, PI: S.S. Henley), National Cancer Institute (NCI) (R44CA139607, PI: S.S. Henley) and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (R43AA014302, PI: S.S. Henley; R43AA013670, PI: S.S. Henley; R43/44AA013351, PI: S.S. Henley; R43/44AA011607, PI: S.S. Henley) under the Small Business Innovation Research (SBIR) program; National Institute on Drug Abuse (5K08DA031245, PI: A.N. Westover).

Funding agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference List

1. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist. Med.* 2002; 21(19):2917–2930.
2. Hamburg MA, Collins FS. The Path to Personalized Medicine. *New England Journal of Medicine.* 2010; 363(4):301–304. [PubMed: 20551152]
3. Petticrew M, Tugwell P, Kristjansson E, Oliver S, Ueffing E, Welch V. Damned if you do, damned if you don't: subgroup analysis and equity. *J Epidemiol. Community Health.* 2012; 66(1):95–98. [PubMed: 21652518]
4. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet.* 2005; 365(9454):176–186.
5. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, az-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC,onso-Coello P, Hassouneh B, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ.* 2012; 344:e1553. [PubMed: 22422832]
6. Wallace ML, Frank E, Kraemer HC. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA Psychiatry.* 2013
7. Akaike, H. Information theory and an extension the maximum likelihood principle. In: Petrov, BN.; Csaki, F., editors. Second international symposium on information theory. Budapest: Akademiai Kiado; 1973. p. 267-281.
8. Burnham, KP.; Anderson, DR. Model Selection and Multi-Model Inference. New York: Springer-Verlag; 2002.
9. Bozdogan H. Akaike's Information Criterion and Recent Developments in Information Complexity. *J Math. Psychol.* 2000; 44(1):62–91. [PubMed: 10733858]
10. Monuteaux MC, Spencer TJ, Faraone SV, Wilson AM, Biederman J. A randomized, placebo-controlled clinical trial of bupropion for the prevention of smoking in children and adolescents with attention-deficit/hyperactivity disorder. *J Clin Psychiatry.* 2007; 68(7):1094–1101. [PubMed: 17685748]
11. Golubchik P, Sever J, Weizman A. Influence of methylphenidate treatment on smoking behavior in adolescent girls with attention-deficit/hyperactivity and borderline personality disorders. *Clin Neuropsychopharmacol.* 2009; 32(5):239–242. [PubMed: 19834989]
12. Winhusen TM, Somoza EC, Brigham GS, Liu DS, Green CA, Covey LS, Croghan IT, Adler LA, Weiss RD, Leimberger JD, Lewis DF, Dorer EM. Impact of attention-deficit/hyperactivity disorder (ADHD) treatment on smoking cessation intervention in ADHD smokers: a randomized, double-blind, placebo-controlled trial. *J Clin Psychiatry.* 2010
13. Pomerleau OF, Downey KK, Stelson FW, Pomerleau CS. Cigarette smoking in adult patients diagnosed with attention deficit hyperactivity disorder. *Journal of Substance Abuse.* 1995; 7(3): 373–378. [PubMed: 8749796]
14. Lambert NM, Hartsough CS. Prospective study of tobacco smoking and substance dependencies among samples of ADHD and non-ADHD participants. *J Learn. Disabil.* 1998; 31(6):533–544. [PubMed: 9813951]
15. Levin ED, Conners CK, Sparrow E, Hinton SC, Erhardt D, Meck WH, Rose JE, March J. Nicotine effects on adults with attention-deficit/hyperactivity disorder. *Psychopharmacology (Berl).* 1996; 123(1):55–63. [PubMed: 8741955]
16. Hughes JR, Keely JP, Niaura RS, Ossip-Klein DJ, Richmond RL, Swan GE. Measures of abstinence in clinical trials: issues and recommendations. *Nicotine. Tob. Res.* 2003; 5(1):13–25. [PubMed: 12745503]
17. Heatherton TF, Kozlowski LT, Frecker RC, Fagerstrom KO. The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br. J Addict.* 1991; 86(9): 1119–1127. [PubMed: 1932883]
18. Dupaul, GJ.; Power, TJ.; Anastopoulos, AD. ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretation. New York: Guilford Press; 1998.

19. Beck, AT.; Steer, RA.; Brown, GK. Beck Depression Inventory-II. San Antonio: The Psychological Corporation, Harcourt Brace & Company; 1996.
20. Beck, AT.; Steer, RA. Beck Anxiety Inventory: Manual. The Psychological Corporation; 1990.
21. Golden, R.; Henley, S.; White, H.; Kashner, T. New Directions in Information Matrix Testing: Eigenspectrum Tests. In: Chen, X.; Swanson, NR., editors. Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis. New York: Springer; 2013. p. 145
22. Hilbe, JM. Logistic Regression Models. New York: Chapman and Hall; 2009.
23. Nunes EV, Covey LS, Brigham G, Hu MC, Levin FR, Somoza EC, Winhusen TM. Treating nicotine dependence by targeting attention-deficit/ hyperactivity disorder (ADHD) with OROS methylphenidate: the role of baseline ADHD severity and treatment response. *J Clin Psychiatry*. 2013; 74(10):983–990. [PubMed: 24229749]
24. White HL. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*. 1982; 50(1): 1–25.
25. Claeskens, G.; Hjort, N. Model Selection and Model Averaging. New York: Cambridge University Press; 2009.
26. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. New York: Springer-Verlag; 2001.
27. Hosmer, DW.; Lemeshow, S. Applied Logistic Regression, Second Edition. New York: John Wiley & Sons, Inc.; 2000.
28. Guideline on the investigation of subgroups in confirmatory clinical trials. European Medicines Agency. 2014 Accessed at http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf.
29. Golden R. Discrepancy Risk Model Selection Test theory for comparing possibly misspecified or nonnested models. *Psychometrika*. 2003; 68(2):229–249.
30. Garside MJ. The Best Subset in Multiple Regression Analysis. 1965:195–200.
31. Furnival GM. All Possible Regressions with Less Computation. *Technometrics*. 1971; 13(2):403–408.
32. Furnival G, Wilson R. Regression by leaps and bounds. 1974:499–511.
33. Hosmer DW, Jovanovic B, Lemeshow S. Best Subsets Logistic Regression. *Biometrics*. 1989; 45(4):1265–1270.
34. Miller, A. Subset Selection in Regression. London: Chapman and Hall; 2002.
35. Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, Schwartz JS. Rethinking Randomized Clinical Trials for Comparative Effectiveness Research: The Need for Transformational Change. *Annals of Internal Medicine*. 2009; 151(3):206–209. [PubMed: 19567619]
36. Conway PH, Clancy C. Comparative-Effectiveness Research GÇö Implications of the Federal Coordinating Council's Report. *New England Journal of Medicine*. 2009; 361(4):328–330. [PubMed: 19567829]
37. Covey LS, Hu MC, Winhusen T, Weissman J, Berlin I, Nunes EV. OROS-methylphenidate or placebo for adult smokers with attention deficit hyperactivity disorder: Racial/ethnic differences. *Drug and Alcohol Dependence*. 2010; 110(1–2):156–159. [PubMed: 20219292]
38. Covey LS, Hu MC, Weissman J, Croghan I, Adler L, Winhusen T. Divergence by ADHD subtype in smoking cessation response to OROS-methylphenidate. *Nicotine. Tob. Res.* 2011; 13(10):1003–1008. [PubMed: 21652734]
39. Covey LS, Hu MC, Green CA, Brigham G, Hurt RD, Adler L, Winhusen T. An exploration of site effects in a multisite trial of OROS-methylphenidate for smokers with attention deficit/ hyperactivity disorder. *Am J Drug Alcohol Abuse*. 2011; 37(5):392–399. [PubMed: 21854282]
40. Berlin I, Hu MC, Covey LS, Winhusen T. Attention-deficit/hyperactivity disorder (ADHD) symptoms, craving to smoke, and tobacco withdrawal symptoms in adult smokers with ADHD. *Drug Alcohol Depend*. 2012; 124(3):268–273. [PubMed: 22364776]
41. Prochaska JJ, Delucchi K, Hall SM. A Meta-Analysis of Smoking Cessation Interventions With Individuals in Substance Abuse Treatment or Recovery. *Journal of Consulting and Clinical Psychology*. 2004; 72(6):1144–1156. [PubMed: 15612860]

42. Buitelaar JK, Kooij JJ, Ramos-Quiroga JA, Dejonckheere J, Casas M, van Oene JC, Schauble B, Trott GE. Predictors of treatment outcome in adults with ADHD treated with OROS(R) methylphenidate. *Prog. Neuropsychopharmacol. Biol. Psychiatry.* 2011; 35(2):554–560. [PubMed: 21185347]
43. Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J.* 2006; 151(2): 257–264. [PubMed: 16442886]
44. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol.* 2001; 54(10): 979–985. [PubMed: 11576808]

Figure 1.
Flow chart depicting analytical steps using a Best Approximating Model approach.

Figure 2. Receiver Operating Characteristic (ROC) Curve for Validated Best Approximating Models of (A) prolonged smoking abstinence (primary outcome), and (B) point prevalence abstinence at Week 10 (secondary outcome).

Table 1

Model Covariates: Participant Demographic and Baseline Characteristics (n=255)

Characteristic	OROS-MPH (n=127)	Placebo (n=128)
Age, years *	38.1 (10.4)	37.5 (9.57)
18–29	32 (25.6%)	36 (28.1%)
30–44	52 (41.6%)	60 (46.9%)
45+	41 (32.8%)	32 (25.0%)
Sex (male)	77 (60.6%)	67 (52.3%)
Race	6 (4.7%)	9 (7.0%)
Non-Hispanic white	98 (76.6%)	104 (81.9%)
Other	30 (23.4%)	23 (18.1%)
Married	51 (40.2%)	35 (27.3%)
Education (years) *	14.4 (2.40)	14.5 (2.41)
<13	32 (25.2%)	32 (25.0%)
13	95 (74.8%)	96 (75.0%)
Employment (full or part-time)	108 (85.0%)	104 (81.3%)
Weight (pounds)	187.28 (44.80)	180.15 (47.65)
Smoking Status		
Fagerstrom score *	5.62 (2.13)	5.45 (2.29)
Number of prior quit attempts *	7.46 (10.3)	6.45 (9.76)
<5	57 (44.9%)	67 (52.3%)
5	70 (55.1%)	61 (47.7%)
Smoking start age, years *	13.9 (3.01)	13.8 (3.17)
<14	52 (40.9%)	63 (49.2%)
14	75 (59.1%)	65 (50.8%)
Lifetime History of Psychiatric Disorders		
Major Depression	41 (32.3%)	46 (35.9%)
Bipolar Disorder	0 (0%)	0 (0%)
Anxiety Disorder	39 (30.7%)	37 (28.9%)
Substance Use Disorder	97 (76.4%)	89 (69.5%)
Baseline Rating Scales		
ADHD Rating Scale total score *	36.0 (7.09)	36.7 (7.49)
36	63 (49.6%)	66 (51.6%)
>36	64 (50.4%)	62 (48.4%)
Beck Depression Inventory score *	9.46 (8.49)	8.70 (7.67)
<14	89 (70.1%)	94 (74.0%)
14	38 (29.9%)	33 (26.0%)
Beck Anxiety Inventory Score *	6.71 (7.07)	5.60 (5.74)
<8	81 (64.3%)	93 (72.7%)
8	45 (35.7%)	35 (27.3%)

Characteristic	OROS-MPH (n=127)	Placebo (n=128)
ADHD Type		
Inattentive	45 (35.4%)	42 (32.8%)
Hyperactive-Impulsive & Combined	82 (64.6%)	85 (66.4%)
Study Site		
1	20 (15.8%)	21 (16.4%)
2	17 (13.4%)	17 (13.3%)
3	32 (25.2%)	31 (24.2%)
4	21 (16.5%)	20 (15.6%)
5	18 (14.2%)	20 (15.6%)
6	19 (15.0%)	19 (14.8%)

* Means and standard deviations.

Table 2

Best Approximating Model of Prolonged Smoking Abstinence

Variable	β Estimates	Standard Error	Wald	P value	Odds Ratio	95% CI
OROS-MPH Treatment (Trt.)	-0.2281	0.6530	0.1220	0.7269	0.7960	(0.221, 2.863)
OROS-MPH Trt. $\times$ Lifetime History of Substance Use Disorder	-1.2937	0.5065	6.5240	0.0107	0.2740	(0.102, 0.740)
OROS-MPH Trt. $\times$ Employment	0.8557	0.5726	2.2330	0.1351	2.3530	(0.766, 7.228)
OROS-MPH Trt. $\times$ ADHD Rating Scale > 36	0.9704	0.4157	5.4480	0.0196	2.6390	(1.168, 5.961)
Sex (male)	0.7132	0.3178	5.0360	0.0248	2.0400	(1.094, 3.804)
Study Site 5	1.1715	0.4309	7.3920	0.0066	3.2270	(1.387, 7.509)
Study Site 2	-0.7867	0.5052	2.4240	0.1195	0.4550	(0.169, 1.226)
Study Site 6	-0.5982	0.4442	1.8130	0.1781	0.5500	(0.230, 1.313)
Study Site 4	0.8729	0.4006	4.7490	0.0293	2.3940	(1.092, 5.249)
Age 18–29	-0.5545	0.3395	2.6670	0.1025	0.5740	(0.295, 1.117)
Fagerstrom Score	-0.1368	0.0655	4.3560	0.0369	0.8720	(0.767, 0.992)
Intercept	0.0391	0.4750	0.0070	0.9345	1.0400	(0.410, 2.638)

GAIC = 325.695, -2LL = 301.864, Max Condition Number = 50.3

GAIC = Generalized Akaike Information Criteria

-2LL = -2 Log Likelihood

Table 3

Best Approximating Model of Smoking Point Prevalence Abstinence in Week 10

Variable	β Estimates	Standard Error	Wald	P value	Odds Ratio	95% CI
OROS-MPH Treatment (Trt.)	0.7569	0.5514	1.884	0.1699	2.132	(0.723, 6.282)
OROS-MPH Trt $\times$ ADHD Rating Scale > 36	0.8106	0.4453	3.314	0.0687	2.249	(0.940, 5.384)
OROS-MPH Trt $\times$ Study site 1	-1.4539	0.7875	3.409	0.0649	0.234	(0.050, 1.094)
OROS-MPH Trt. $\times$ Age 18-29 years	-1.4667	0.6076	5.827	0.0158	0.231	(0.070, 0.759)
OROS-MPH Trt $\times$ Lifetime History of Substance Use Disorder	-0.6404	0.4918	1.696	0.1928	0.527	(0.201, 1.382)
Beck Anxiety Inventory Score 8	-0.4081	0.3297	1.532	0.2158	0.665	(0.348, 1.269)
Employment (full- or part-time)	0.5416	0.4519	1.436	0.2307	1.719	(0.709, 4.168)
Study Site 5	2.0309	0.4803	17.877	0.0000	7.621	(2.973, 19.537)
Study Site 3	0.0108	0.4084	0.001	0.9790	1.011	(0.454, 2.251)
Study Site 4	0.9472	0.4377	4.683	0.0305	2.578	(1.093, 6.080)
Study Site 1	1.2006	0.5328	5.078	0.0242	3.322	(1.169, 9.439)
Intercept	-1.5383	0.5405	8.100	0.0044	0.215	(0.074, 0.619)

GAIC = 317.988, -2LL = 293.325, Max Condition Number = 11.1

GAIC = Generalized Akaike Information Criteria

-2LL = -2 Log Likelihood

Table 4

Hypothesis-Driven Subgroup Analysis for Each Individual BAM Covariate Contrasted with Previously Published Analyses and BAM Results.

Covariates	Hypothesis-Driven Approach		Planned/Published Analyses		BAM results	
	Prolonged Smoking Abstinence	Point Prevalence Abstinence	P-value for interaction *	Planned Analysis	Significant in Published Analysis? †	Significant? †
Demographic						
Age (18–29, 30–44, 45 years)	0.5189	0.2587			No	Yes
Sex (male)	0.2004	0.6685	✓		No	No
Race (Non-Hispanic White)	0.4070	0.4472	✓		No	No
Education (0–12, 13 years)	0.4920	0.4763			No	No
Married	0.4886	0.3999			No	No
Employment (Part/Full-Time) [§]	0.2901	0.9710			No	No
Baseline Measures						
Weight (pounds)	0.3373	0.2263			No	No
No. of smoking quit attempts (0–4, 5)	0.5137	0.8333			No	No
Smoking start age (0–13, 14 years)	0.6083	0.3496	✓		No	No
Fagerstrom score [‡]	0.6128	0.7523	✓		No	No
Major depression (lifetime)	0.9069	0.8630	✓ ^{††}		No	No
Bipolar disorder (lifetime)	— [¶]	— [¶]	✓ ^{††}		No	No
Anxiety disorder (lifetime)	0.9295	0.7305	✓ ^{††}		No	No
Substance use disorder (lifetime)	0.0978	0.6611	✓		Yes	No
ADHD rating scale score (cutpoints 0–36, >36)	0.0018	0.0147	✓		No ^{‡(37)} / Yes ⁽²³⁾	No
Beck depression inventory score (cutpoints 0–13, 14)	0.1792	0.9045	✓ ^{‡‡}		No	No
Beck anxiety inventory score (cutpoints 0–7, 8)	0.1940	0.7790	✓ ^{‡‡}		No	No
ADHD subtype (Hyperactive/Impulsive + Combined types, Inattentive type)	0.0343	0.1924	✓		Yes ⁽³⁵⁾	No

Study Characteristic

Covariates	Hypothesis-Driven Approach		Planned/Published Analyses		BAM results		
	Prolonged Smoking Abstinence	Point Prevalence Abstinence	P-value for interaction *	Planned Analysis	Significant in Published Analysis? †	Prolonged Smoking Abstinence	Point Prevalence Abstinence
	P-value for interaction *	P-value for interaction *				Significant? ‡	Significant? ‡
Study Sites (by six individual sites)	0.2472	0.1111				No	No
Study Site (by clinic type) **	0.4535	0.9780	✓		No ⁽³⁶⁾		

* P-values derived from logistic regression analyses. Maximum likelihood estimates were used in all cases except for covariates with more than two categories where Type 3 effect was used.

† Significance defined as $p < 0.05$.

‡ Continuous covariates were analyzed around the mean value, i.e. those subjects higher than the mean value versus those lower than the mean value.

[§] Employment status was dichotomized as employed full-/part-time or unemployed.

[¶] There were no observations of lifetime bipolar disorder.

^{***} Each of the six study sites were designated as one of the following categories: mental health clinic, ADHD clinic, or tobacco dependence clinic.

^{††} Lifetime major depression, bipolar disorder, and anxiety disorder were combined as a single measure to indicate psychiatric comorbidity for a planned analysis, and were not analyzed separately.

^{‡‡} ADHD rating scale, Beck depression and anxiety inventory scores were analyzed as continuous variables (i.e. no cutpoints) for planned analyses, in these cases.

Table 5

Comparison of Sun et al. Criteria for Subgroup Effect Credibility versus BAM Characteristics

Sun et al. Criteria for Assessing Credibility of Subgroup Effect	Characteristics of Best Approximating Model (BAM) Analysis
<i>Design</i>	
Was the subgroup variable a baseline characteristic?	BAM may include only baseline characteristics.
Was the subgroup variable a stratification factor at randomization?	BAM may be used whether or not covariates were used as stratification factors during randomization.
Was the subgroup hypothesis specified a priori?	Candidate covariates for BAM analysis are specified a priori. BAM hypothesizes that there is a best model that best represents the data generating process.
Was the subgroup analysis one of a small number of subgroup hypotheses tested (5)?	BAM avoids multiple testing by carrying out a pre-specified model selection algorithm using an information-theoretic approach.
<i>Analysis</i>	
Was the test of interaction significant (interaction $P < 0.05$)?	BAM analyzes all covariates for treatment interaction.
Was the significant interaction effect independent, if there were multiple significant interactions?	BAM tests all covariate-treatment interactions for independence. Higher-order interactions (to an investigator-specified level) may be tested as well.
<i>Context</i>	
Was the direction of subgroup effect correctly prespecified?	BAM does not prespecify a direction of effect.
Was the subgroup effect consistent with evidence from previous related studies?	The results of a BAM analysis may be compared to evidence from previous studies.
Was the subgroup effect consistent across related outcomes?	BAM may be performed on multiple related outcomes.
Was there any indirect evidence to support the apparent subgroup effect—for example, biological rationale, laboratory tests, animal studies?	BAM may be compared with indirect evidence