Addressing Methodologic Challenges and Minimizing Threats to Validity in Synthesizing Findings from Individual-Level Data Across Longitudinal Randomized Trials

Ahnalee Brincks[1] · Samantha Montag[2,3] · George W. Howe[4] · Shi Huang[5] · Juned Siddique[2,3] · Soyeon Ahn[6] · Irwin N. Sandler[7] · Hilda Pantin[8] · C. Hendricks Brown[2,3]

Published online: 22 April 2017
© Society for Prevention Research 2017

Abstract Integrative Data Analysis (IDA) encompasses a collection of methods for data synthesis that pools participant-level data across multiple studies. Compared with single-study analyses, IDA provides larger sample sizes, better representation of participant characteristics, and often increased statistical power. Many of the methods currently available for IDA have focused on examining developmental changes using longitudinal observational studies employing different measures across time and study. However, IDA can also be useful in synthesizing across multiple randomized clinical trials to improve our understanding of the comprehensive effectiveness of interventions, as well as mediators and moderators of those effects. The pooling of data from randomized clinical trials presents a number of methodological challenges, and we discuss ways to examine potential threats to internal and external validity. Using as an illustration a synthesis of 19 randomized clinical trials on the prevention of adolescent depression, we articulate IDA methods that can be used to minimize threats to internal validity, including (1) heterogeneity in the outcome measures across trials, (2) heterogeneity in the follow-up assessments across trials, (3) heterogeneity in the sample characteristics across trials, (4) heterogeneity in the comparison conditions across trials, and (5) heterogeneity in the impact trajectories. We also demonstrate a technique for minimizing threats to external validity in synthesis analysis that may result from non-availability of some trial datasets. The proposed methods rely heavily on latent variable modeling extensions of the latent growth curve model, as well as missing data procedures. The goal is to provide strategies for researchers considering IDA.

Keywords Integrative data analysis · Synthesis methodology · Participant-level meta-analysis · Harmonization

Introduction

Integrative Data Analysis (IDA; Curran and Hussong 2009) is a framework for statistical analysis on a single dataset resulting from the pooling of individual participant data across multiple studies. Unlike meta-analysis, which combines parameter estimates from previously completed analyses on the separate studies, IDA estimates new parameters using the raw data combined across multiple

Electronic supplementary material The online version of this article (doi:10.1007/s11121-017-0769-1) contains supplementary material, which is available to authorized users.

✉ C. Hendricks Brown
hendricks.brown@northwestern.edu

[1] Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI, USA

[2] Department of Psychiatry and Behavioral Sciences, Feinberg School of Medicine, Northwestern University, 750 N. Lake Shore Dr, 10th Floor, Chicago, IL 60611, USA

[3] Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

[4] Department of Psychology, George Washington University, Washington, DC, USA

[5] Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

[6] School of Education and Human Development, University of Miami, Miami, FL, USA

[7] Department of Psychology, Arizona State University, Tempe, AZ, USA

[8] Department of Public Health Sciences, Miller School of Medicine, University of Miami, Miami, FL, USA

studies. In this paper, we discuss analytic challenges and solutions to using IDA to examine intervention effects across multiple trials. Despite its challenges, IDA enables researchers to examine intervention response across numerous samples incorporating a broader range of risk and protective factors. This pooling of individual-level data across multiple trials also has a distinct advantage of enhancing statistical power allowing for more complex statistical models (Brown et al. 2013; Dagne et al. 2016).

IDA presents a number of methodological challenges for data preparation and analytic modeling. The first of these is measure harmonization. Even in a study combining trials targeting the same outcome, say depressive symptoms, measurement of outcomes, as well as baseline covariates may differ across trials. A second challenge is heterogeneity in the populations represented across trials. Trials generally differ in population characteristics of race/ethnicity, geographic region, community socioeconomic status, as well as community cultures and political histories. Prevention trials often differ dramatically in the level of baseline risk, with some trials being universal while others are selective or indicated. A third area of challenge includes heterogeneity in study characteristics. These differences include characteristics unique to the protocol such as the timing of assessments or the mode and target of delivery. This also includes the type of intervention and the number of intervention conditions in the trial.

Some of these challenges have been addressed in the literature. For instance, Item Response Theory (IRT) has been put forth as one method for handling measure harmonization. The method relies on bridging items that are common across different measures and determines dimensionality of the underlying construct, tests for differential item functioning based on subsets of the samples (e.g., gender and age), and finally creates a scale score (Bauer and Hussong 2009; Curran and Hussong 2009; Curran et al. 2008). This IRT approach could well be extended to multiple dimensions using a bifactor model (Gibbons and Hedeker 1992), although we are unaware of the use of the IRT and bifactor modeling in synthesizing intervention effects across multiple trials. A second method involves bridging measures rather than bridging items. Here, at least one study includes assessments on more than one measure (Siddique et al. 2015). This approach enumerates all measures used in any of the trials and formally treats the measures not used in a trial as missing data. Multiple imputation can then be used to fill in the gaps with imputed data based on the correlations across measures (Siddique et al. 2016) or a factor model can be imposed on the data as presented in Brown et al. (2016).

Concerns resulting from heterogeneity across trials have also been addressed in the literature. When IDA is undertaken with a small number of studies, or a group of studies that are not considered representative of an entire population of such studies, an indicator of trial membership can be included as a fixed covariate in fixed effect modeling. With a large number of trials, multilevel or random effects modeling can be used as it is in meta-analysis (see Hedges and Vevea 1998 for discussion of their relative merits and Hussong et al. 2013 for specific discussion of their use in IDA). Finally, multilevel mixture meta-analysis modeling can be used to identify different clusters of trials as well as outcomes where effects are distinct (Brown et al. 2008).

While solutions to these individual methodologic challenges of IDA analysis have been addressed individually in the literature, these challenges compound quickly when working with data from multiple, longitudinal, randomized clinical trials where follow-up assessment periods, outcome measures, sample characteristics, and control or intervention groups are likely to differ between trials. Even across trials targeting the same outcome, the intervention approaches may differ significantly in terms of delivery target (e.g., delivered to individuals vs. delivered to groups), theoretical framework (e.g., based in interpersonal therapy vs. cognitive behavioral therapy), and number of intervention arms (e.g., control compared with single intervention, control compared with multiple interventions, and comparison of two active interventions). All of these challenges raise questions about the internal validity of the synthesis study. External validity, the concern that selection of those trials included in a synthesis may differ from the universe of available trials, is also an important question that needs to be addressed.

The methodologic aims of this manuscript are to articulate methods that address generic questions of internal and external validity in an IDA synthesis study that pools individual participant data across multiple, longitudinal randomized clinical trials. To address challenges to internal validity, we articulate methods and the underlying assumptions used to handle (1) different outcome measures used in different trials, (2) different follow-up assessments employed by the different trials, (3) differences in sample characteristics across trials, (4) combining results across trials containing different comparison conditions, (5) modeling variation in impact trajectories, relying heavily on extensions to latent growth curve modeling, and (6) assessing selection bias threats to external validity based on non-availability of some trial datasets. We address these methodological aims through the example of Brown et al. (2016), an IDA of multiple adolescent depression prevention trials aimed at clarifying the overall, and differential, impact of depression-preventive interventions on trajectories of depressive symptoms.

The data from Brown et al. (2016) consist of a combined individual-level dataset of 19 randomized controlled preventive trials targeting depression, general mental health, or problem behaviors among 5292 adolescents. The goal of the original synthesis study was to better understand the impact of preventive intervention on adolescent internalizing symptoms, including depression as well as anxiety, withdrawal, and related symptoms. The original study also uncovered sources of variation in the impact of the intervention as a result of the robust, diverse sample created by pooling across the multiple trials. Findings indicated an overall beneficial impact on 2-year depressive symptom trajectories, with better outcomes for adolescents involved in trials that were depression-focused, employed interventions based on cognitive behavioral therapy or interpersonal therapy, and/or were delivered directly to the adolescent rather than a parent/guardian. The analyses presented significant challenges due to variability in outcome measures, timing of follow-up assessments, participant samples, and intervention design. The trials and findings are described in detail in Brown et al. (2016), and in this paper, we focus on the underlying methodologic approaches used therein.

In this paper, we first provide a brief overview of these 19 trials and their variations in measures, assessment, intervention, and sample characteristics. As many of the challenges faced in analysis can be viewed from a missing data perspective, we discuss this view next. Third, we present six challenges, their potential for affecting internal or external validity, and what analytic strategies we have used to reduce these threats. Finally, we discuss general lessons that are applicable to other synthesis projects and methodologic approaches.

Variability Across Trials in the Collaborative Data Synthesis for Adolescent Depression Trials

We discuss general challenges and ways to minimize threats to validity for synthesizing findings from an analysis of individual-level data from multiple trials; we illustrate solutions based on pooling of individual-level data across the 19 prevention trials (Perrino et al. 2013). All these 19 trials began between 1991 and 2007; they randomized and repeatedly assessed 5210 adolescents meeting our age restriction of 11 to 18 years old. There were some basic similarities across these trials (see Suppl Table A, to be published as Table 1 in Brown et al. 2016). All tested at least one preventive intervention against a control condition in a randomized trial successfully conducted under a rigorous protocol. However, they differed in major ways regarding participant selection criteria, measures of adolescent internalizing behavior, follow-up schedules, and the interventions themselves. For this prevention synthesis project, the trials were deliberately chosen to include variations on these characteristics, so that we could examine an overall effect of prevention interventions on internalizing symptoms and, where possible, to identify moderators and mediators of that effect, capitalizing on the enhanced power from the large sample size and increased variability afforded by the pooled dataset (Brown et al. 2013).

Measures Variability

A general issue encountered in synthesizing the effects of interventions across trials on a behavioral primary outcome is almost never measured with the same instrument across all studies. Interventionists interested in measuring adolescent depression have a wide range of measures to choose from, and this presents a significant challenge in synthesis work. In our study, eight measures of adolescent depression were identified from the trials: four self-report measures (Youth Self Report—Anxiety/Depression; Youth Self Report—Withdrawal/Depression, Center for Epidemiologic Studies—Depression Scale (CESD), and the Children's Depression Inventory (CDI)), three parent-report measures (Revised Behavior Problem Checklist—Anxiety/Withdrawal; Child Behavior Checklist (CBC)—Anxiety/Depression; Child Behavior Checklist—Withdrawal/Depression), and one clinician rating (Children's Depression Rating Scale). Descriptions of the measures are in Brown et al. (2016). We report elsewhere (Howe et al. 2016) on psychometric analyses of a combined individual-level dataset of 123 items available from baseline depression measures in 16 of these studies. These item response theory analyses provided support for a single common factor, along with some variations by raters, particularly for clinician ratings. In addition, there was evidence for some differential item functioning when comparing boys and girls, although these effects appear relatively small compared with the global depression and rater effects. In evaluating the impact of these trials, our general analytic approaches to these diverse measures was to consider them as distinct, observed indicators of an underlying latent variable representing internalizing symptoms (Brown et al. this issue), as we describe below. This unobserved, latent variable is modeled as the source of the association between the diverse observed measures of internalizing symptoms (i.e., "effect indicators"; Bollen and Lennox 1991).

Assessment Variability

Another universal regarding trials is that follow-up assessment schedules vary across trials. Across our 19 trials, the available longitudinal data ranges from 6 months to 15 years (see Suppl Fig. A, to be published as Fig. 2 in Brown et al. 2016). Most trials included 4 to 6 time points and our approach grouped follow-up times into clusters so the data could be analyzed as a panel study. This neglected exact observation times but allowed us to estimate a correlation structure at each time interval. Studying the pattern of available data points, we clustered follow-up assessment periods into six time clusters up to 24 months post-baseline. Assessments beyond 24 months were dropped because of the small number of trials having such length of follow-up. Note that all time blocks

Table 1 Comparison of methods

Models for synthesis of legacy trials	Key elements	Infrastructure to support synthesis	Methodology strengths	Methodology challenges
Meta-analysis	Assemble summary statistics from published reports	Standard methods and tools available (e.g., Cochrane Collaboration)	Methodology well established	Results are limited to findings that have been published or completed; difficult to harmonize findings across different instruments; severely limited to main effect or subgroup analyses; subject to ecological fallacies
Repository of individual-level data from trials	Structured data system to document and share all available trial data	Federal research mandates to submit all future trials and availability of a shared database	Can conduct mediator and moderator analyses; ability to link individuals across trials; defined structure for data, documentation, and experiments	Potential misinterpretation data and low-quality control involving analyses by individuals not connected to the trial; legacy trials not included
Partnering to conduct integrative data analysis	Partnering of trial directors to answer complex synthesis questions	Technology for assembling, harmonizing, and analyzing data from multiple trials	Ability to conduct sophisticated analyses with comparable data across studies	Challenging to obtain individual-level data from all available trials
Parallel data analysis	Have individual sites conduct analyses on their own data based on a standard protocol, then combine findings at a central site	Supplemental funding of individual trials for new follow-up data and analysis	Does not require sharing of individual-level data; ability to conduct sophisticated analyses with comparable data across studies	Provide technical assistance for producing comparable analyses that can be combined

National Database for Clinical Trials related to Mental Illness (NDCT) (http://ndct.nimh.nih.gov/). Advancing Science Through Collaborative Data Sharing and Synthesis (Perrino et al. 2013)

had some data available across all these 19 trials, and all trials included follow-up assessments through the 14–18-month time block.

Intervention Variability

There are very few examples of exact replication of interventions in behavioral research, at least in part because a funding and scientific emphasis on innovativeness over repetitiveness. The 19 trials included in our analysis certainly differ in both substance and scope. Although all of the trials were prevention trials, they varied in important ways. Only 9 of the 19 trials specifically targeted the prevention of depression. The other ten trials targeted other important outcomes (general mental health, externalizing, substance use, and high-risk sexual behavior), and included measures of depressive symptoms as part of their protocol. Eleven trials utilized two intervention arms, seven trials utilized three intervention arms, and one trial utilized four arms. We categorized the active interventions as focusing on cognitive behavioral therapy (CBT), interpersonal therapy (IPT), or parenting skills development. Eleven trials tested active control arms such as another evidence based intervention (e.g., IPT, CBT, or the active intervention of interest without a key component such as parent groups) and the remaining eight utilized control arms such as bibliotherapy.

Sample Variability

Prevention trials in a synthesis often include widely different populations that range from universal, selective, to indicated. In our example, the 19 trials differed in terms of their inclusion/exclusion criteria with some trials targeting adolescents who had been in trouble with authority, other trials focusing on adolescents who had recently experienced a loss or family change, and others employing a universal approach. These differences in intervention target also resulted in samples that differed significantly not only in terms of important comorbidities like externalizing but also on demographic variables, specifically income and parent education (see Suppl Table A, to be published as Table 2 in Brown et al. 2016). Sample sizes for the individual trials ranged from 41 to 697 adolescents, and eight trials also included parents who attended either separate or conjoint intervention sessions. Race/ethnicity of the participants was reported for all but four trials and included White/Caucasian, Hispanic (Non-White), Black/African-American, Asian, Hawaiian/Pacific Islander, Native American/Alaskan Native, and Other/Multiple. Females comprised between 42 and 85% of the included participants with the exception of one trial that included solely females, and the age of participants ranged from 11 to 18 years. Adolescents were evaluated with a variety of psychosocial, development, and depression measures, with the CDI, CESD, and CBC being the most common.

Considering Synthesis as a Missing Data Problem

In any synthesis of multiple longitudinal trials, missing data can be categorized into five distinct types. The first includes data that were missing within a trial due to attrition or incomplete response; a type of missingness that occurs in virtually all longitudinal trials. For example, a single participant may have missed the scheduled 6-month follow-up interview for his or her trial. Or the proportion of missing items left unanswered by a subject at a point in time may have exceeded 20% of the number of items for that measure, a routine cutoff that we used to classify the composite score as missing. The second form is missingness as a result of measure selection. For example, each trial team selected from a wide range of depressive symptom measures; those measures not selected for a particular trial can be considered missing data. Third, the different follow-up schedules in each of the 19 trials can be considered as creating incomplete panel data. Fourth, we can consider each subject in a two-arm trial having two potentially observable measures, one being the set of longitudinal responses if assigned to the active intervention and the other being the set of longitudinal responses if assigned to the control condition. Which of these outcomes is missing depends on the random assignment to condition. Finally, the fifth type of missing data is truncation. We only observe data from the trials whose datasets were shared with us; any other data that could be available from different trials are unknown. Our approach to truncation is very different from that used to handle the other types of general censored data; we discuss this last situation under external validation.

Ignorability of Details Regarding Why Data Are Missing and Inferential Approach

Our general analytic approach relies on full information maximum likelihood (FIML) to handle all types of missing data under a missing at random (MAR) assumption. MAR assumes that missingness is unrelated to outcomes once observed data are taken into account. FIML requires this assumption in order to guarantee unbiased estimates when including data on predictor variables when some outcomes are missing. FIML conditions on observed covariates and maximizes this conditional likelihood after averaging over any missing data, treating that portion of the likelihood involving missing data as irrelevant, or ignorable for making inferences. Methods for analyzing missingness of an individual's datum at a point in time on a particular measure in a single randomized trial (the first of the types of missing data listed above) has been heavily investigated by statisticians. While no analytic method can be expected to produce accurate inferences under all possible missing data mechanisms when there are large amounts of missing data, approaches that use FIML are known to be highly robust in such situations (Brown 1990; Lavori et al. 2008; Siddique et al. 2008).

Now consider the second, third, and fourth types of missing data listed above, involving trials having different measures, different follow-up times, and missingness on conditions to which they were not randomly assigned (e.g., active intervention if assigned to control). We argue that for this set of well-conducted trials and each of these situations, the data can reasonably be considered to follow mechanisms classified as missing at random (Rubin 1976). Missing at random is a technical term describing an important condition whereby the exact reasons for why data are missing have no effect on the inferences one can draw from the data. Thus, the data are called ignorably missing. For missing at random to occur, the missing data must have the same distribution as the observed data once we condition on observed covariates. Under conditions of missing at random, the method of full information maximum likelihood is fully efficient in making inferences as long as the underlying distribution is modeled accurately, and we include all necessary covariates (Rubin 1976).

It is not immediately obvious that we can reasonably ignore missing data mechanisms if all the trials have different follow-up designs with different measures. Different trials have different reasons for missingness (e.g., more universal trials may involve less follow-up and more self-report measures), but whatever the reason, it should be identical for intervention and control groups within each trial as assessment schedules would be the same across conditions and assessors should be blind to condition as well. Thus, the second through fourth types of missingness should be ignorable. However, trial then becomes a critical variable to account for in our analysis. Because we know which subjects belong to which trials, trial is an observed covariate for everyone, and analyses that formally include models where each trial is permitted to have a distinct pattern of growth, the inclusion of trial as a fixed effect can then lead to appropriate inferences. Including trial as a random effect, rather than a fixed effect, is also not likely to produce much bias in our overall impact analysis of the difference between intervention and control, but such analyses may possibly hide important variation as they involve averaging over the entire set of trials. We also note for completeness that latent variables are missing for everyone and therefore missing at random (Brown 1983).

The estimation of these models also requires that the pattern of observed variables across the trials is sufficient to identify all the parameters. For example, if one particular measure

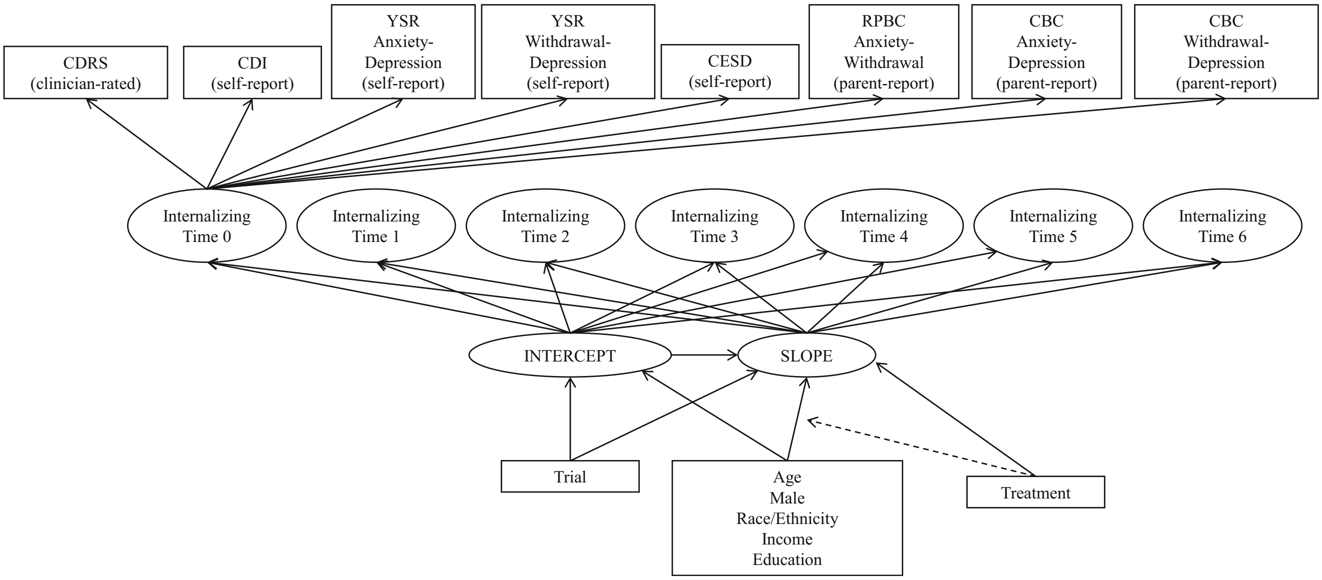

Fig. 1 Second-order latent growth curve model. Though not pictured, the observed depression measures are indicators of the latent internalizing construct at each time point

is only observed for a single trial, there is no information available to assess its correlation with other measures (Dagne et al. 2016; Howe et al. 2016).

A final important note is about trial selection for inclusion into a synthesis study of this kind. As indicated in Fig. 1a from Brown et al. (2016; see Suppl Fig. B), we obtained participant-level data from 76% of the trials that we attempted to include, but many other trials do exist. We discuss this issue of truncation under external validity below.

Challenge No. 1: Heterogeneity in Outcome Measures

Handling heterogeneity in the primary outcomes is the first major task to address in any IDA. In our example, the outcome measures for adolescent internalizing symptoms varied across the 19 trials, and this presented a significant challenge in harmonizing across the studies. We identified eight measures of adolescent depression that were most common across the 19 trials in order to best capture a common construct of internalizing across all trials.

Extensive procedures were used to ensure equivalent coding across trials. Whenever possible, we coded data based on original items rather than precomputed constructs in each of the trials. This allowed us to standardize how summary scores were computed across all trials. For example, individual research teams may have differed in how they handled missing items in the computation of a summary score. By working with item-level data, we standardized this across trials so that a summary score was considered missing if less than 80% of the items were completed by a participant, which is a common construct level decision used by researchers.

Variants in measures abound, as instruments are shortened or extended by individual researchers depending on the time available for a survey or by their unique interests. By design, some of the trials in our example did not measure the entire set of available items on a particular instrument, or used a related, custom version of an instrument. We treated these shortened and custom measures as “surrogate” measures and incorporated them into our analysis design. In one example, a trial administered a custom set of items closely related to the CDI on 77% of the participants, and the full CDI on the remaining 23%. The overlap of items between this custom measure and the CDI allowed us to regress the summary score for the custom measure on the CDI in our model and thereby infer CDI scores on these participants. In another trial, a shortened version of the CESD was used instead of the full measure, and we used a similar regression approach to include these participants. The regression coefficients for this relationship were calculated based on all the trials having full CESD items, as the short form scores were computable. Also, we fixed the regression coefficients to be the same across time panels and trials, then used full information maximum likelihood to account for these surrogates in the analysis.

In any IDA, one needs to arrive at a conceptual definition of the primary outcome of interest. In this example, our primary outcome variable was an unobserved latent variable of depressive symptoms identified by the eight measures of depressive and internalizing symptoms represented across the 19 trials. We estimated this latent variable model using only baseline values first to assess for fit and found it was moderately good given the large number of parameters involved

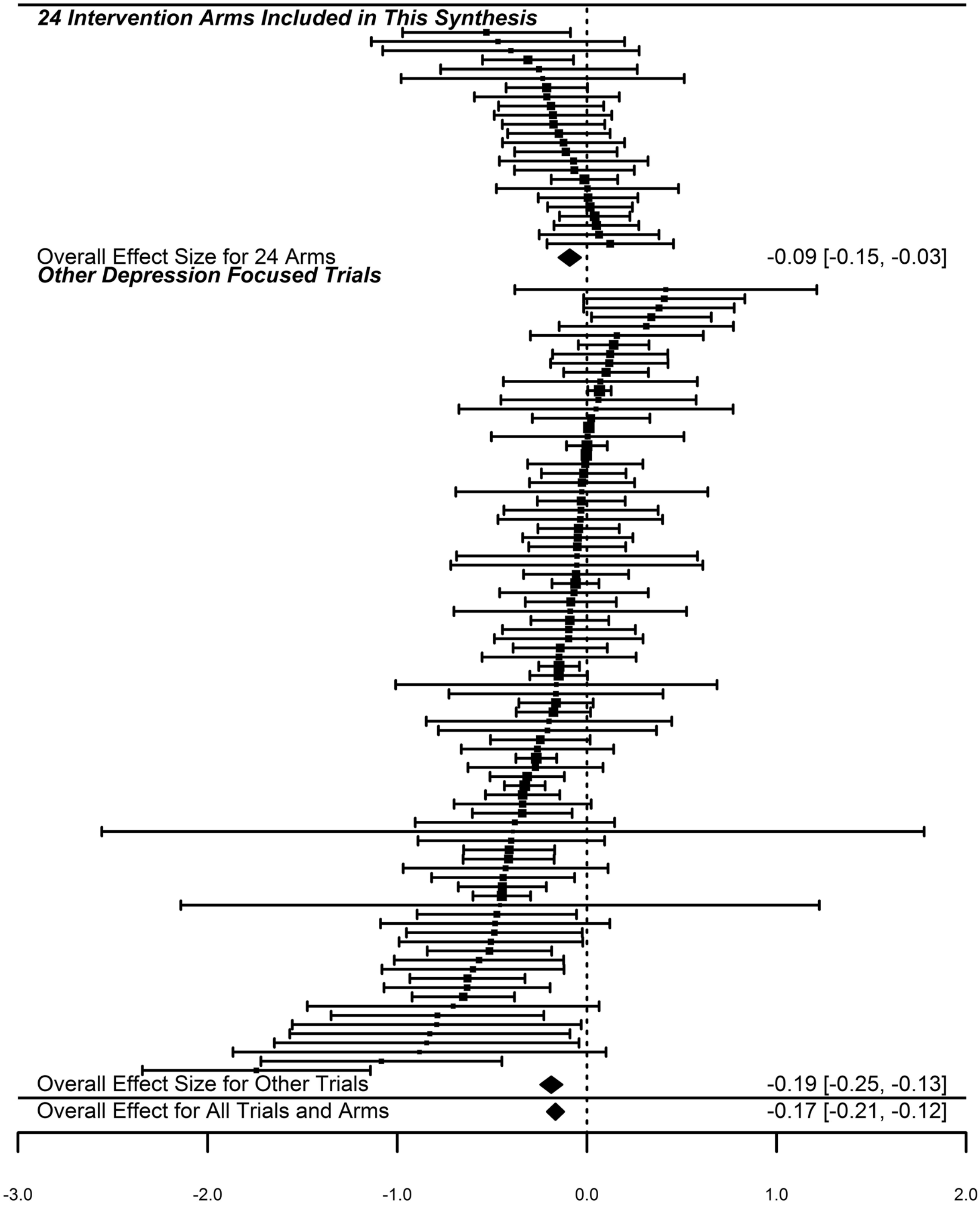

Fig. 2 Forest plot of effect sizes for our 19 trials and the greater universe of trials. *Asterisk*, effect size where negative effect represents beneficial intervention. *Double asterisk*, these trials did not specifically target depression with their intervention

(CFI = 0.88, RMSEA = 0.018). In the longitudinal model described below, we used the same latent variable structure at each time point and fixed the path coefficients between the observed measure of depressive symptoms and the latent variable to be equal across time. This ensures stability or consistency of the latent construct over time (Hancock et al. 2001). The assumption of loading invariance over time is commonly considered necessary for growth models, although there is evidence that modest violations of this assumption do not bias results in many cases (Edwards and Wirth 2012). Further, work is clearly necessary to develop methods for testing invariance in complex multilevel models of the sort we employed. The regression of surrogate measures, as described above, were also part of this structure (Fig. 1). This latent variable approach enabled us to measure a construct of "depressive symptoms" for every participant in the 19 trials at each time interval, regardless of the number of observed measures of depressive symptoms available for that particular participant. As discussed above, the missing data can be thought of as "planned missingness" resulting from trial design.

The requirement for internal validity is that the factor analysis model posits a single underlying latent construct of depressive symptoms that covers self-reports, parent reports, and clinical reports and assumes the same measurement error structure across trials and time. As indicated, the model fit for latent depressive symptoms was adequate, indicating general support for the one-dimensional structure. While a formal test of equivalence across all trials and times would likely find some significant variations, we feel reassured by the stability of the loadings that we found across the many different models that we fit. We note that for our analyses in this paper, we have relied on depressive symptom data at the scale level. For future follow-up analyses, we plan to use the items themselves in a more complex item response analysis (Howe et al. 2016).

Challenge No. 2: Heterogeneity in Follow-up Assessment Periods

As times of follow-up will differ across trials, some decisions need to be made regarding which of two approaches should be used to accommodate these variations. One approach is to model data using each individual's own time points; the second is to cluster similar time points into batches and analyze as a panel study. Both approaches have advantages and disadvantages; the former approach is taken in Siddique et al. (2016) and discussed more fully therein. In Brown et al. (2016), we took a multivariate, panel data approach for longitudinal data, as opposed to the alternative hierarchical linear model or mixed effect modeling approach (Raudenbush and Bryk 2002) because it provided greater capacity to test models within a structural equation modeling framework using latent growth curve modeling. We first placed each trial's follow-up times into six time blocks which collected assessment points that were reasonably close together, thus creating a panel design. The resulting time blocks were as follows: baseline (time 0 (T0)), which was measured for everyone, 7 days to less than 2 months (time 1 (T1)), 2 months to less than 5.5 months (time 2 (T2)), 5.5 months to less than 9 months (time 3 (T3), 9 months to less than 14 months (time 4 (T4)), 14 months to less than 24 months (time 5 (T5)), and 24 months (time 6 (T6)). Only a few trials had measures at time 1 (7 days to 2 months), and all of those that did also had measures at time 2 (2 to 3 months). Figure 2 from Brown et al. (2016; see Suppl Fig. A) identifies the number of youth at baseline enrolled into the trial within ages 11–18 and shows the availability of follow-up data for each of the trials at each time period up to 24 months post-intervention. Note that all time blocks had some data available across the 19 trials, with 6 of the trials measured up to 24 months, and all trials had at least one assessment through the14- to <24-month block.

The major analytic challenge in this example, as well as with other IDA for randomized trials, involves estimating latent growth trajectories across time and outcome measure and relate that to overall and specific intervention effects. In our example, we did this by modeling a single latent construct for depressive symptoms for the *i*th subject in the *k*th trial (η_{ikt}) at each time point (t). Across the seven assessment periods, we constrained the loadings and intercepts on the measurement model for depressive symptoms to be equal and allowed the factor variances to change with time, controlling for covariates. Specifically, the *j*th depression measure assessed at time t for the *i*th subject in the *k*th trial, $Y_{ij\mathrm{kt}}$, is an indicator of the underlying latent variable, η_{ikt} representing the underlying latent construct of depression

$$Y_{\mathrm{ijkt}} = \mu_j + \lambda_j \times \eta_{ikt} + \varepsilon_{\mathrm{ijkt}}, i = 1, \ldots, N_k; j = 1, \ldots, 8; k = 1, \ldots, 19, t = 0, \ldots, 6 \tag{1}$$

Here, μ_j and λ_j are means and factor loadings for the *j*th measure, and ε_{ijkt} is an error term, considered to have a normal distribution with zero mean and unique error $\sigma_i{}^2$. In these models, trials were treated as fixed effects.

The latent constructs were used as indicators of a latent growth model to estimate change in depression across time. Specifically, we allowed for a general transformation of the time axis to transform time ($f(t)$) that would linearize the effect of the intervention. We allowed our modeling of the growth over time, i.e., "slope" to capture any linear as well as nonlinear pattern across time points. This is especially important in prevention as prevention effects may diminish or even reverse over time, and such patterns would not be detected if we forced the pattern to be linear. Since linearity is specified in a second-order growth model with loadings on the slope latent variable equal to the time point where each of the six panels were obtained (i.e., 0 for baseline, 6- for 24-month outcome), nonlinearity was accounted for by allowing the second-level factor loadings of the growth model to be estimated by the data (see Suppl Fig. C, to be published as Fig. 3 in Brown et al. (2016)). Latent variable growth modeling methods can easily handle different follow-up times across trials, but still needs to represent the overall pattern of growth adequately across time and trial. Our approach permits the data to inform how the pattern evolved over time and is quite general, enabling recognition of linear or non-linear type of growth.

Specifically, our second-order model at the level of the individual specified how these underlying latent depression variables relate to individual-level growth curves specified by their intercepts α_{ik} and slopes β_{ik} on this transformed time scale:

$$\eta_{ikt} = \alpha_{ik} + \beta_{ik} f(t) + \varepsilon_{ikt}, i = 1, \ldots, N_k; k = 1, \ldots, 19, t = 0, \ldots, 6 \tag{2}$$

Growth models that employ latent factors in this way have been referred to as "second-order latent growth models" (Hancock et al. 2001), "curve of factors models" (McArdle 1988), and "latent variable longitudinal curve models" (Tisak and Meredith 1990).

Our second-order latent growth model included these two latent variables: baseline internalizing (α_{ik} or intercept) and a latent variable for linear change on internalizing (β_{ik} or slope). In our structural equation modeling, we controlled on individual-level demographic variables of age, gender, race/ethnicity, family income, and parent's educational attainment for both the second-level intercept and slope. We also adjusted the intercept and the slope for trial as a categorical factor. This introduction of fixed effects for trial was used instead of two random effects for intercept and slope because the number of trials was too small to estimate these variances and covariances with sufficient precision given the few measures used in each trial. We were also concerned that single random effects may not represent trial-level heterogeneity sufficiently well. Because baseline levels of internalizing may well influence the trajectory of internalizing, we regressed the slope on the intercept to control for baseline internalizing. A test of the intervention effect on each (two-arm) trial was based on the impact on slopes of the indicator of intervention status, $Z_{ik} = 0$ for control and 1 for intervention, after adjusting for a vector of other covariates $\mathbf{X}_{ik}$.

$$\beta_{ik} = \tau_k + \pi \times \alpha_{ik} + \theta_k \times Z_{ik} + \boldsymbol{\delta}^{\ddot{E}}\mathbf{X}_{\mathrm{ik}} + \varepsilon_{ik}, i = 1, \ldots, N_k;; k = 1, \ldots, 19 \tag{3}$$

Here. the trial-specific intervention versus control effects are given by θ_k and individual-level errors ε_{ik}. The overall effect of the intervention versus control can be obtained using the model,

$$\theta_k = \theta + \varepsilon_k, k = 1, \ldots, 19 \tag{4}$$

where θ is an overall mean, ε_k represents variation at the trial level, and testing of intervention impact is based on H_0: $\theta = 0$ against a two-sided alternative.

A test of moderation of the intervention effect on slope trajectory by baseline level of depressive symptoms was conducted by testing whether the regression coefficients of the latent slope on the latent intercept were different for intervention versus control. The same approach was used to test for other moderator effects at the individual or trial level.

A primary assumption for internal validity in this approach is that variation in change over time within each cluster of follow-up time points is negligible. Since the follow-up assessments that are grouped together are relatively close in time, we are comfortable making this assumption, but recognize that we are unable to fully test for its accuracy. When we examine the estimated trajectory of time (see Suppl Fig. C, to be published as Fig. 3 in Brown et al. (2016)), the amount of change between 9 and 14 months (time period 4) has a very small rate of change, suggesting that this assumption is valid.

Challenge No. 3: Sample Heterogeneity

When attempting to combine effects across diverse trials, an important issue for generalizability is the controlling of between-study differences in the samples. Though there was some overlap due to common prevention goals in our example, each of the intervention studies had unique inclusion/exclusion criteria, leading to differences in presence or level of internalizing or externalizing symptoms at baseline. There were also major differences due to socioeconomic status, with some studies having a more disadvantaged sample than others. Race/ethnicity was another source of between-study heterogeneity, with some trials focused exclusively on participants from a particular ethnic or racial group, and other studies inclusive of participants from multiple racial/ethnic groups. A final important source of between-trial heterogeneity in the samples involved the intervention itself. Some interventions were based in treatment modalities such as Cognitive Behavioral Therapy or Interpersonal Therapy. Other interventions had strong family components with a significant amount of time invested with parents of the target adolescent.

Our first global approach to controlling for between-study differences in the sample was to include trial membership as a covariate. For our study of 19 trials, this meant including 18 dummy-coded variables as covariates on the baseline level of internalizing (intercept) and the trajectory of internalizing across time (slope). Such modeling allows for baseline differences in internalizing across trials, which clearly is supported by the data, and potential shifts in the course of symptoms in control groups by trial. The particular choice of dummy coding, i.e., which trial is used as a contrast, has no effect on the coefficient of overall intervention impact that is our primary interest. By using this approach, we acknowledge that this collection of 19 trials is not necessarily a random selection from a broader population of similar studies. Instead, we incorporate trial membership as a fixed effect within the analytic model which effectively removes variability due to differences between trials (Curran 2009). A key assumption for internal validity in this approach is that the effects of the covariates do not vary across trial. That is, the tests of interactions between trial, covariate, and intervention are not statistically significant.

Challenge No. 4: Multiple Intervention Arms/Arm-Level Analyses

In the previous section, we described our approach to examining an overall impact across all trials. This method is appropriate as long as there is at least one active intervention and a comparison condition in each trial. However, more can be done when there are multiple intervention arms versus a single control, or even when there are intervention arms tested against one another without a control within the same trial, as one would find in a comparative effectiveness trial. We illustrate how we would approach these issues through our specific example. In our example, each of the 19 trials included one or two active intervention arms in the trial to compare against one control condition. We were interested in the characteristics of the intervention that may moderate treatment effects, such as type of intervention (CBT, IPT, and family-based), recipient of the intervention (child, parent, and both conjointly) and whether the trial targeted depression as an outcome. All the active arms in the trial were coded based on these characteristics; all codes were checked for accuracy by the trial principal investigators. This provided us an opportunity to assess whether intervention impact varied by type of intervention or modality.

We attempted several types of analysis, the most general being two-level growth modeling in Mplus (i.e., three-level mixed effects modeling involving time, person, and arm of trial). Unfortunately, these analyses did not converge due to the modest number of trial arms (24) when fit with two correlated random effects for intercept and slope. As an alternate approach, we estimated within-trial intervention effects for each active arm against control using the second-order growth model with slope regressed on each trial (adjusting for intercept and individual-level covariates) and extracted these adjusted empirical Bayes estimates and their standard errors into a separate dataset. There were 6 of the 19 trials with two active intervention arms, and these were compared with the same control condition, resulting in the estimation of 24 effects comparing the active intervention versus control. In a single trial with two active arms and one control arm, the two empirical Bayes estimates are dependent and required that we account for such non-independence when analyzing arm-level covariates. For trials with two active arms, we used orthogonal transformations to rotate these effect sizes as well as the covariates and then revise their respective standard errors to uncorrelated estimates while retaining the same multilevel mean and variance-covariance structure. These orthogonal transformations (Cholesky decompositions) were based on the eigenvalue-eigenvector decomposition of the level-one variance covariance matrix that preserved the total variance at level one as well as the variance across the trials. Details can be obtained from the last author.

Challenge No. 5: Examining Variation in Pathways to Assess Differential Impact

As this project examined variation in impact, we included analytic modeling that allowed subjects in the study to vary both qualitatively and quantitatively in their response to intervention. Following our earlier method work (Muthén and Brown 2009; Muthén et al. 2002, 2010; Wang et al. 2005), we used growth mixture models to assess such variations in response. Specifically, an underlying mixture distribution provided each cluster with its own overall trajectory shape and degree of variation, and each cluster had its own impact measure as well. Because these data all came from randomized trials, the clusters were parameterized so that the parameters for the baseline distributions within each trial were the same for intervention as they were for control (e.g, same-class probability, intercept mean, and variance). We also held the measurement error the same across clusters (Eq. 1). To permit the intervention effects to vary across cluster, we allowed the underlying time transform to linearity, $f(t)$ in Eq. 2 to vary. To assess potentially different intervention effects, we allowed parameters that captured differences in the distribution of the slope conditional on the intercept to vary by cluster as well (e.g., π, θ_k, and its residual variance). Our main finding was that the beneficial effect of these preventive interventions occurred among those who started with an elevated, but subclinical level of symptoms (Brown et al. 2016).

Challenge No. 6: Assessing Trial Selection Bias

The primary threat to external validity in a study of this kind is related to the question of selection bias. That is, is the sample of trials for which data are available representative of the full universe of trial data in this substantive area? With respect to external validity, we are interested in understanding how we can relate the findings from these analyses to other prevention trials in the literature that are not included in this synthesis. Additionally, from the perspective of our research questions, we would be less confident in our conclusions if we found the included trials had stronger impact than did other prevention trials whose individual data we did not obtain. Thus, we looked for an existing report on trials that were comparable with ours on a common outcome measure. Sandler et al. (2014) had recently conducted an overview of published meta-analyses on trials that focused on preventing depression in youth. We used this overview to locate 85 distinct trials and coded their effect sizes post-intervention on the primary measure of depressive symptoms used by each of these trials. Effect sizes for our trials that were reviewed by one or more of these meta-analyses were extracted, but for the 16 of our 19 trials that were not included in this meta-analytic overview, we computed effect sizes (ES) at 6 months from our data for each intervention condition versus control.

Figure 2 provides the forest plot of the effect sizes of the greater universe of trials compared with the 24 intervention arms for the 19 trials included in this synthesis. Using random effects models, we found effect sizes to be significantly different from zero for both our comparisons and those of the other trials. The effect size for our trials showed a less beneficial effect than the greater universe of trials (ES = −0.09 compared with ES = −0.19, respectively). Despite a small overlap in the 95% confidence interval for our trials (−0.15, −0.03) and the greater universe of trials (−0.25, −0.13), these effect sizes are statistically different (z = 2.26, p = 0.024). Because the overall effect size for the selected trials was closer to zero than the effect size for the universe of trials, we concluded that our approach has not selected the most significant trials, which provides a more serious bias in inference than that of the other direction.

To address representativeness of the trials in our synthesis, we also compared the standard deviations in these two sets of studies, finding twice as much spread in the 86 other depression prevention trials compared with ours. This is not surprising as the other prevention trials contained interventions tested in widely different settings (e.g., schools) and tested on different ages. Other than this smaller variance for our trials, we did not find any other concerns that our select group of trials differed from the larger population of trials.

We also extrapolated our growth model findings for our 24 comparisons to the remaining preventive trials by regressing the difference in intervention versus control slopes on the ESs. Such a procedure is commonly used in model-based survey estimation when there is no formal random sampling selection as is true in our study (Royall and Cumberland 1981). This extrapolated effect based on our model-based approach using post-intervention ESs resulted in virtually no change in our inference because the correlation between ESs and our adjusted slope differences in intervention versus control was near zero after accounting for measurement error at the trial level.

Discussion

The pooling of individual participant data across multiple, longitudinal, randomized trials is rich with methodological challenges that have implications regarding internal and external validity. Some of these validity issues are shared with meta-analysis. Methodologically, these analyses were difficult to carry out. Not only did trials vary by time of follow-up, they used different measures to assess internalizing symptoms. Many of the methodological approaches employed here have been used in isolation rather than combination (Curran et al. 2008). The present work builds upon this literature by showing how these multiple challenges can be addressed simultaneously using a large number of diverse randomized clinical trials. The statistical methods we used in this paper, factor analysis modeling to address measurement error in internalizing symptoms, different types of growth modeling to connect outcomes across time, full information maximum likelihood to handle missing data, and growth mixture modeling to assess variation in response, were able to work together to provide inferences that painted a clear picture of impact. This was not without some cost. The maximization algorithms that handled missing data for some models took days to converge. It would have been possible to compute more complex models in this study with over 5000 longitudinal observations had there not been such variations in measures and time points. Specifically, we were unable to achieve convergence on any growth mixture models involving three or more classes. Also, these variations limited our ability to check our model. Thus, tests on the stability of the underlying factor model over time were limited by the trials' different instruments and follow-up times.

Within IDA, there are also alternate approaches to the latent variable methods we employed. For instance, missing data, including those data missing as a result of trial design, can be imputed using multiple imputation methods, although such methods have their own limits, as discussed in Siddique et al. (2016). Siddique et al. (2016) conducted growth models that used individual time points rather than pooling them together. They found similar conclusions to those in Brown et al. (2016), but these methods encountered more problems. Specifically, they were forced to throw away individual cases and trials because of the sparseness of measures across these studies. Similarly, item-level analyses could be implemented using procedures similar to those in the original IDA studies (Curran et al. 2008). One perspective that is well represented by such work is to identify items that have little differential item functioning over time and across subpopulations. Separating such items may provide clearer pictures of impact than those methods that we used that relied on summaries of items. There are major computational challenges to overcome with the use of IRT-type models with such large, longitudinal datasets, and moderation effects are anticipated to be challenging to evaluate.

As research moves further in the direction of synthesizing participant-level data across trials, the use of common, well-established measures across trials would certainly simplify the analysis problems and provide additional opportunities for assessing goodness of fit. The perspective of using common measures across studies has a time-honored tradition in science and is one of the major policy changes now under way at the National Institutes of Health. Indeed, the PhenX Toolkit (Barch et al. 2016; Hendershot et al. 2015; Pathak et al. 2011) includes measures that now must be used by US federally funded researchers by default, and thereby lessen the harmonization burden of combining data from different sources (Barch et al. 2016; McCarty et al. 2014; Pan et al. 2012). While common measures are important in moving science forward, as attested to the challenges our example gave, we do have three caveats. First, use of the same instrument by different raters can produce quite variable results as Siddique et al. (2011) have shown. Secondly, there are some advantages to employing two or more respected measures of important constructs, particularly target outcomes. Their inclusion would strengthen the latent variable structures presented here by providing more overlap in the measures across studies. Third, the PhenX and other approaches to standardizing assessments have the advantage of enforcing the use of the same first-level measures and therefore can reduce some of the harmonization problems. However, another approach, which we believe will be used more often is computerized adaptive testing or CAT (Gibbons et al. 2012). The algorithmic nature of CAT selects the items for each person to be most informative, based on responses to previous items. Thus, individuals receive different items until a predetermined level of measure reliability is achieved. Based on a large pool of items and psychometric studies that account for multiple dimensions, CAT procedures can be highly efficient and readily used in synthesis studies. As the science improves our understanding of when best to measure long-range effects on internalizing symptoms and other important outcomes, it may be possible to establish more common follow-up assessment schedules that would also strengthen these synthesis analyses by making it possible to model time more precisely and limiting the reliance on missing data methods. Such common schedules may become even more important as decade-long preventive effects are being discerned (Sandler et al. 2014).

Integrative data analysis, whether it is based on a repository of trials or is built through a partnership, is one of three analytic approaches to synthesizing the effects of an intervention across randomized clinical trials. The others include meta-analysis and parallel data analysis (Brown et al. 2013), and in Table 1, we compare the strengths and challenges of each method. Meta-analytic techniques provide an established method for synthesizing across the summary statistics (e.g., effect sizes) reported in completed trials. It is comparatively inexpensive, but as noted previously, this approach is limited by assumptions about the comparability of outcome measures, the potential for selection bias when relying on published trials, and the general inability to assess for mediators and moderators, particularly at the level of the individual. Parallel data analysis involves combining analytic summaries based on a common analytic protocol that is carried out separately by each individual research team. Thus, parallel data analysis moves closer to IDA by estimating similar statistical models across each individual trial separately, and then summarizing across the resulting estimates. While parallel data analysis is rarely performed, it can be a highly efficient method to use, especially when combined with strategically funded funding for longer-term follow-up or assessment of measures not in the original trial (Brown et al. 2007, 2013).

We have seen the value of IDA in addressing variation in impact through measured as well as unmeasured covariates that interact with intervention. Procedures for sharing of data through partnership have been identified (Perrino et al. 2013) and their potential for addressing scientific equity for minorities and other populations has been noted elsewhere (Perrino et al. 2015). This partnering does take considerable time to develop, and the requirement of data sharing may lead to highly select samples of trials. However, these disadvantages can be countered by investigating selection bias against non-shared trials as we have done in this paper and the benefit of close feedback by the original research teams who have deep tacit knowledge of their respective studies (Brown et al. 2013). The existence of a repository of studies, such as NIMH's National Database for Clinical Trials Related to Mental Illness (NDCT) and NIDA's National Addiction and HIV Archive Program (NADHAP), are likely to have a profound effect on the sharing of future trials.

There are a number of limitations inherent in this study. Even with the large number of trials in this synthesis, we recognize that there are many other prevention programs that have been tested in trials that we have not included. Although we examined effect sizes for trials not captured in this synthesis and found they were on average reporting more beneficial outcomes, we cannot rule out the possibility that the trials we analyzed differed in other important ways. With the inclusion of more trials, we would certainly enhance the overall findings presented in Brown et al. (2016). Our use of extrapolated empirical Bayes estimates to address dependency inherent in multiple-arm trials is a less ideal method than using multilevel modeling, limited in this trial by the small number of comparisons across our 19 trials. This paper was limited to one IDA synthesis, and we did not include any simulation studies of how these advanced methods would work together. Thus, we cannot guarantee that our experiences will generalize. With respect to measurement limitations, the latent construct of depressive symptoms is based on the assumption that the observed measures are capturing the same symptom levels for members of subgroups within and across trials. Additionally, the use of measures that employ different reporters (youth, parent, and clinician) also presents the possibility of differential item functioning (Bauer et al. 2013) that is not accommodated in the measure-level model used in this study. It will be important to pursue future item-level psychometric work that investigates the possibility for, and potential implications of, such effects in synthesis data. A final, important, limitation was our limited ability to test moderated effects across important classifications of interventions, including how they are delivered and to whom. This occurred because of the large degree of overlap (e.g., most cognitive based preventive interventions were delivered to higher income and more educated populations). This particular limitation is thus not limited to IDA.

Acknowledgments We gratefully acknowledge the National Institute of Mental Health Collaborative Data Synthesis for Adolescent Depression Trials Study Team, composed of our many colleagues who generously provided their data to be used in this study, obtained access to key datasets, reviewed coding decisions, provided substantive or methodologic recommendations, and reviewed this manuscript. We also thank NIMH for their support through grant number R01-MH040859 (Collaborative Data Synthesis for Adolescent Depression Trials, Brown PI) and NCATS for their support through grant number UL1-TR000460 (Miami Clinical and Translational Science Institute, Szapocznik PI), as well as the following grants that financially supported the trials: NIMH R01-MH048696 (Beardslee), NIMH R01-MH064717 (Beardslee & Gladstone), NIMH R01-MH064503 (Brent), NIMH R03-MH048118 (Clarke), NIDA R01-DA07031 (Dishion), NIDA R01-DA13773 (Dishion), NIMH R01-MH064735 (Garber), WT Grant Foundation 961730 (Garber), NIMH R01-MH064707 (Gonzales), NIDA R01-DA017462 (Pantin), NIDA R01-DA025894 (Prado), CDC U01-PS000671 (Prado), NIMH P30-MH068685 (Sandler), NIMH R01-MH049155 (Sandler & Ayers), NIMH R01-MH052270 (Seligman), NIDA R01-DA007029 (Spoth), NIDA R01-DA018374 (Stormshak), NIMH MH063042 (Szapocznik), NIMH K08-MH072918 (Van Voorhees), NIMH K23-MH071320 (Young), and NARSAD Young Investigator Grant (Young). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies nor that of our collaborators who provided access to their data.

Compliance with Ethical Standards

Conflict of Interest CHB was supported as a consultant on one of these projects (New Beginnings) and received funding support on the Familias Unidas trials as did Brincks, Huang, and Pantin who developed and directed much of research on this intervention. Sandler developed and directed much of the work on two trials in this synthesis.

Ethics Approval This article does not contain any studies with animals performed by any of the authors. This project involving sharing of data was review by two IRBs. Use of deidentified data in this synthesis project was approved by the University of Miami and Northwestern IRBs, and all institutions signed data use agreements with Northwestern.

Informed Consent All trials included in this synthesis were approved by their respective institutional review boards. Informed consent was obtained from all individual participants included in these respective studies.

References

Barch, D. M., Gotlib, I. H., Bilder, R. M., Pine, D. S., Smoller, J. W., Brown, C. H., …, Farber, G. K. (2016). Common measures for national institute of mental health funded research. *Biological Psychiatry, 79*, e91–e96. doi:10.1016/j.biopsych.2015.07.006.

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods, 14*, 101–125. doi:10.1037/a0015583.

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods, 18*, 475–493. doi:10.1037/a0032475.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin, 110*, 305–314.

Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika, 48*, 269–291.

Brown, C. H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics, 46*, 143–155.

Brown, C. H., Wyman, P. A., Brinales, J. M., & Gibbons, R. D. (2007). The role of randomized trials in testing interventions for the prevention of youth suicide. *International Review of Psychiatry, 19*, 617–631. doi:10.1080/09540260701797779.

Brown, C. H., Wang, W., & Sandler, I. (2008). Examining how context changes intervention impact: the use of effect sizes in multilevel meta-analysis. *Child Development Perspectives, 2*, 198–205. doi:10.1111/j.1750-8606.2008.00065.x.

Brown, C. H., Sloboda, Z., Faggiano, F., Teasdale, B., Keller, F., Burhart, G., …, Prevention Science and Methodology Group (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. *Prevention Science, 14*, 144–156. doi:10.1007/s11121-011-0207-8.

Brown, C. H., Brincks, A., Huang, S., Perrino, T., Cruden, G., Pantin, H., …, Sandler, I. (2016). Two-year impact of prevention programs on adolescent depression: An integrative data analysis approach. *Prevention Science*, 1–21.

Curran, P. J. (2009). The seemingly quixotic pursuit of a cumulative psychological science: Introduction to the special issue. *Psychological Methods, 14*, 77–80. doi:10.1037/a0015972.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*, 81–100. doi:10.1037/a0015914.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., & Zucker, R. A. (2008). Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. *Developmental Psychology, 44*, 365–380. doi:10.1037/0012-1649.44.2.365.

Dagne, G., Brown, C. H., Howe, G., Kellam, S., & Liu, L. (2016). Testing moderation in network meta-analysis with individual participant data. *Statistics in Medicine, 35*, 2485–2502. doi:10.1002/sim.6883.

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: a longitudinal example. In J. R. Harring, G. R. Hancock, J. R. Harring, & G. R. Hancock (Eds.), *Advances in longitudinal methods in the social and behavioral sciences* (pp. 289–311). Charlotte: IAP Information Age Publishing.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry, 69*, 1104–1112. doi:10.1001/archgenpsychiatry.2012.14.

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling, 8*, 470–489. doi:10.1207/S15328007SEM0803_7.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504. doi:10.1037/1082-989X.3.4.486.

Hendershot, T., Pan, H., Haines, J., Harlan, W. R., Marazita, M. L., McCarty, C. A., …, Hamilton, C. M. (2015). Using the PhenX Toolkit to add standard measures to a study. Current Protocols in

Human Genetics, 86, 1.21.21–17. doi:10.1002/0471142905.hg0121s86.

Howe, G. W., Dagne, G., Brincks, A., & Beardslee, W. (2016). Evaluating construct equivalence and harmonizing measurement of adolescent depression when synthesizing results across multiple studies. *Submitted for publicaton*.

Hussong, A. M., Curran, P. J., & Bauer, D. J. (2013). Integrative data analysis in clinical psychology research. *Annual Review of Clinical Psychology, 9*, 61–89. doi:10.1146/annurev-clinpsy-050212-185522.

Lavori, P. W., Brown, C. H., Duan, N., Gibbons, R. D., & Greenhouse, J. (2008). Missing data in longitudinal clinical trials. Part A. Design and conceptual issues. *Psychiatric Annals, 38*, 784–792.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York: Plenum Press.

McCarty, C. A., Huggins, W., Aiello, A. E., Bilder, R. M., Hariri, A., Jernigan, T. L., …, Zeng, Y. (2014). PhenX RISING: Real world implementation and sharing of PhenX measures. BMC Medical Genomics, 7, 16. doi:10.1186/1755-8794-7-16.

Muthén, B. O., & Brown, C. H. (2009). Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling. *Statistics in Medicine, 28*, 3363–3395. doi:10.1002/sim.3721.

Muthén, B. O., Brown, C. H., Masyn, K., Jo, B., Khoo, S. T., Yang, C. C., …, Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. Biostatistics, 3, 459–475. doi:10.1093/biostatistics/3.4.459.

Muthén, B. O., Brown, C. H., Leuchter, A., & Hunter, A. (2010). General approaches to analysis of course: Applying growth mixture modeling to randomized trials of depression medication. In P. E. Shrout, K. M. Keyes, & K. Ornstein (Eds.), *Causality and psychopathology: finding the determinants of disorders and their cures* (pp. 159–178). Washington: American Psychiatric Publishing.

Pan, H., Tryka, K. A., Vreeman, D. J., Huggins, W., Phillips, M. J., Mehta, J. P., …, Ramos, E. M. (2012). Using PhenX measures to identify opportunities for cross-study analysis. Human Mutation, 33, 849–857. doi: 10.1002/humu.22074.

Pathak, J., Pan, H., Wang, J., Kashyap, S., Schad, P. A., Hamilton, C. M., …, Chute, C. G. (2011). Evaluating phenotypic data elements for genetics and epidemiological research: Experiences from the eMERGE and PhenX Network Projects. AMIA Summits on Translational Science Proceedings, 2011, 41–45.

Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., Shern, D., Brown, C. H.(2013). Advancing science through collaborative data sharing and synthesis. *Perspectives on Psychological Science, 8*(4), 433–444. doi:10.1177/1745691613491579.

Perrino, T., Beardslee, W., Bernal, G., Brincks, A., Cruden, G., Howe, G., …, Brown, C. H. (2015). Toward scientific equity for the prevention of depression and depressive symptoms in vulnerable youth. Prevention Science, *16*, 642–651. doi:10.1007/s11121-014-0518-7.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. Thousand Oaks: Sage Publications.

Royall, R. M., & Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association, 76*, 66–77.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.

Sandler, I. N., Wolchik, S. A., Cruden, G., Mahrer, N. E., Ahn, S., Brincks, A., & Brown, C. H. (2014). Overview of meta-analyses of the prevention of mental health, substance use, and conduct problems. *Annual Review of Clinical Psychology, 10*, 243–273. doi:10.1146/annurev-clinpsy-050212-185524.

Siddique, J., Brown, C. H., Hedeker, D., Duan, N., Gibbons, R. D., Miranda, J., & Lavori, P. W. (2008). Missing data in longitudinal trials—part B, analytic issues. *Psychiatric Annals, 38*, 793–801.

Siddique, J., Crespi, C. M., Gibbons, R. D., & Green, B. L. (2011). Using latent variable modeling and multiple imputation to calibrate rater bias in diagnosis assessment. *Statistics in Medicine, 30*, 160–174. doi:10.1002/sim.4109.

Siddique, J., Reiter, J. P., Brincks, A., Gibbons, R. D., Crespi, C. M., & Brown, C. H. (2015). Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. *Statistics in Medicine, 34*, 3399–3414. doi:10.1002/sim.6562.

Siddique, J., de Chavez, P. J., Howe, G., Cruden, G., & Brown, C. H. (2016). Limitations in using multiple imputation to harmonize individual participant data for meta-analysis. *Prevention Science*, 1–14.

Tisak, J., & Meredith, W. (1990). Descriptive and associative developmental models. In A. Von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. II, pp. 387–406). San Diego: Academic Press.

Wang, C.-P., Brown, C. H., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association, 100*, 1054–1076. doi:10.1198/016214505000000501.