2021-10-01

Meaningful associations in the adolescent brain cognitive development study

Anthony Steven Dick
Florida International University

Et al.

Meaningful associations in the adolescent brain cognitive development study

Anthony Steven Dick [a], Daniel A. Lopez [b], Ashley L. Watts [c], Steven Heeringa [d], Chase Reuter [e],
Hauke Bartsch [f], Chun Chieh Fan [g], David N. Kennedy [h], Clare Palmer [i], Andrew Marshall [j],
Frank Haist [k], Samuel Hawes [a], Thomas E. Nichols [l], Deanna M. Barch [m], Terry L. Jernigan [h],
Hugh Garavan [n], Steven Grant [o], Vani Pariyadath [o], Elizabeth Hoffman [p], Michael Neale [q],
Elizabeth A. Stuart [r], Martin P. Paulus [s], Kenneth J. Sher [c], Wesley K. Thompson [e,g,*]

[a] Department of Psychology and Center for Children and Families, Florida International University, Miami, FL, United States
[b] Division of Epidemiology, Department of Public Health Sciences, University of Rochester Medical Center, Rochester, NY 14642, United States
[c] Department of Psychology, University of Missouri, MO, United States
[d] Institute for Social Research, University of Michigan, Ann Arbor, MI 48109, United States
[e] Herbert Wertheim School of Public Health and Human Longevity Science, University of California, San Diego, La Jolla, CA 92093, United States
[f] Mohn Medical Imaging and Visualization Center, Department of Radiology, Haukeland University Hospital, Bergen, Norway
[g] Population Neuroscience and Genetics Lab, University of California, San Diego, La Jolla, CA 92093, United States
[h] Department of Psychiatry, University of Massachusetts Medical School, MA United States, 01604
[i] Center for Human Development, University of California, San Diego, La Jolla, CA 92093, United States
[j] Children's Hospital Los Angeles, and the Department of Pediatrics, University of Southern California, Los Angeles, CA, United States
[k] Department of Radiology, University of California, San Diego, La Jolla, CA 92093, United States
[l] Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[m] Departments of Psychological & Brain Sciences, Psychiatry and Radiology, Washington University, St. Louis, MO 63130, United States
[n] Department of Psychiatry, University of Vermont, Burlington, VT, 05405, United States
[o] Behavioral and Cognitive Neuroscience Branch, Division of Neuroscience and Behavior, National Institute on Drug Abuse, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, United States
[p] National Institute on Drug Abuse, National Institutes of Health, Department of Heatlh and Human Services, Bethesda, MD, United States
[q] Department of Psychiatry, Virginia Commonwealth University, Richmond, VA 23298, United States
[r] Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States
[s] Laureate Institute for Brain Research, Tulsa, OK, United States

ARTICLE INFO

ABSTRACT

The Adolescent Brain Cognitive Development (ABCD) Study is the largest single-cohort prospective longitudinal study of neurodevelopment and children's health in the United States. A cohort of $n = 11,880$ children aged 9–10 years (and their parents/guardians) were recruited across 22 sites and are being followed with in-person visits on an annual basis for at least 10 years. The study approximates the US population on several key sociodemographic variables, including sex, race, ethnicity, household income, and parental education. Data collected include assessments of health, mental health, substance use, culture and environment and neurocognition, as well as geocoded exposures, structural and functional magnetic resonance imaging (MRI), and whole-genome genotyping. Here, we describe the ABCD Study aims and design, as well as issues surrounding estimation of meaningful associations using its data, including population inferences, hypothesis testing, power and precision, control of covariates, interpretation of associations, and recommended best practices for reproducible research, analytical procedures and reporting of results.

1. Introduction

The Adolescent Brain Cognitive Development[SM] (ABCD) Study is the largest single-cohort long-term longitudinal study of neurodevelop-

ment and child and adolescent health in the United States. The study was conceived and initiated by the United States' National Institutes of Health (NIH), with funding beginning on September 30, 2015. The ABCD Study® collects observational data to characterize US population trait distributions and to assess how biological, psychological, and environmental factors (including interpersonal, institutional, cultural, and physical environments) can relate to how individuals live and develop

in today's society. From the outset, the NIH and ABCD scientific investigators were motivated to develop a baseline sample that reflected the sociodemographic variation present in the US population of 9–10 year-old children, and to follow them longitudinally through adolescence and into early adulthood.

The ABCD Study was designed to address some of the most important public health questions facing today's children and adolescents (Volkow et al., 2018). These questions include identifying factors leading to the initiation and consumption patterns of psychoactive substances, substance-related problems, and substance use disorders as well as their subsequent impact on the brain, neurocognition, health, and mental health over the course of adolescence and into early adulthood. More broadly, a large epidemiologically informed longitudinal study beginning in childhood and continuing on through early adulthood will provide a wealth of unique data on normative development, as well as environmental and biological factors associated with variation in developmental trajectories. This broader perspective has led to the involvement of multiple NIH Institutes that are stakeholders in the range of health outcomes targeted in the ABCD design. (Information regarding funding agencies, recruitment sites, investigators, and project organization can be obtained at https://abcdstudy.org).

Population representativeness, or more precisely, absence of uncorrected selection bias in the subject pool, is important in achieving external validity, i.e., the ability to generalize specific results of the study to US society at large. As described below, the ABCD Study attempted to match the diverse US population of 9–10 year-old children on key demographic characteristics. However, even with a largely representative sample, failure to account for key confounders can affect internal validity, i.e., the degree to which observed associations accurately reflect the effects of underlying causal mechanisms. Moreover, it is crucial that the study collects a rich array of variables that may act as moderators or mediators, including biological and environmental variables, in order to aid in identifying potentially causal pathways of interest, to quantify individualized risk for (or resilience to) poor outcomes, and to inform public policy decisions. External and internal validity also depend on assessing the impact of random and systematic measurement error, implementing analytical methods that incorporate relevant aspects of study design, and emphasizing robust and replicable estimation of associations.

The ABCD Study primary aims are given in the Supplementary Materials (SM) Section S.1. We describe the study design and outline analytic strategies to address the primary study aims, including worked examples, with emphasis on approaches that incorporate relevant aspects of study design (Section 2: Study Design; Section 3: Population Weighting). We emphasize the impact of sample size on the precision of association estimates and thoughtful control of covariates in the context of the large-scale population neuroscience data produced by the ABCD Study (Section 4: Hypothesis Testing and Power; Section 5: Effect Sizes; Section 6: Control and Confounding Variables), and we briefly outline state-of-the-field recommendations for promoting reproducible science (Section SM.5) and best practices for statistical analyses and reporting of results using the ABCD Study data (Section SM.6). For readability, more technical subject matter is also largely left to SM sections.

2. Study design

The ABCD Study is a prospective longitudinal cohort study of US children born between 2006 and 2008. A total cohort of $n = 11,880$ children aged 9–10 years at baseline (and their parents/guardians) was recruited from 22 sites (with one site no longer active) and are being followed for at least ten years. Eligible children were recruited from the household populations in defined catchment areas for each of the study sites during the roughly two-year period beginning September 2016 and ending in October 2018.

2.1. Recruitment

Within study sites, consenting parents and assenting children were primarily recruited through a probability sample of public and private schools augmented to a smaller extent by special recruitment through summer camp programs and community volunteers. ABCD employed a probability sampling strategy to identify schools within the catchment areas as the primary method for contacting and recruiting eligible children and their parents. This method has been used in other large national studies (e.g., Monitoring the Future (Bachman et al., 2011); the Add Health Study (Chantala and Tabor, 1999); the National Comorbidity Replication-Adolescent Supplement (Conway et al., 2016); and the National Education Longitudinal Studies (Ingels et al., 1990)). Twins at four "twin-hub" sites were recruited from birth registries (see (Garavan et al., 2018; Iacono et al., 2017) for participant recruitment details). A minority of participants were recruited through non-school-based community outreach and word-of-mouth referrals.

2.2. Inclusion criteria

Across recruitment sites, inclusion criteria consisted of being in the required age range and able to provide informed consent (parents) and assent (child). Exclusions were minimal and were limited to lack of English language proficiency in the children, the presence of severe sensory, intellectual, medical or neurological issues that would impact the validity of collected data or the child's ability to comply with the protocol, and contraindications to MRI scanning (Garavan et al., 2018). Parents must be fluent in either English or Spanish.

2.3. Measures

Measures collected in the ABCD Study include a neurocognitive battery (Luciana et al., 2018; Thompson et al., 2019), mental and physical health assessments (Barch et al., 2018), measures of culture and environment (Zucker et al., 2018), biospecimens (Uban et al., 2018), structural and functional brain imaging (Casey et al., 2018; Hagler et al., 2018), geolocation-based environmental exposure data, wearables and mobile technology (Bagot et al., 2018), and whole genome genotyping (Loughnan et al., 2020). Many of these measures are collected at in-person annual visits, with brain imaging collected at baseline and at every other year going forward. A limited number of assessments are collected in semi-annual telephone interviews between in-person visits. Data are publicly released on an annual basis through the NIMH Data Archive (NDA, https://nda.nih.gov/abcd). Fig. 1 graphically displays the measures that have been collected as part of the ABCD NDA 3.0. Release. Fig. 2 depicts the planned data collection and release schedule over the initial 10 years of the study.

2.4. Sociodemographics

ABCD sample baseline demographics (from NDA Release 2.0.1, which contains data from $n = 11,879$ subjects) are presented in Table 1, along with a comparison to the corresponding statistics from the American Community Survey (ACS). The ACS is a large probability sample survey of US households conducted annually by the US Bureau of Census and provides a benchmark for selected demographic and socioeconomic characteristics of US children aged 9–10 years. The 2011–2015 ACS Public Use Microsample (PUMS) file provides data on over 8000,000 sample US households. Included in this five-year national sample of households are 376,370 individual observations for children aged 9–10 and their households.

With some minor differences, the unweighted distributions for the ABCD baseline sample closely match the ACS-based national estimates for demographic characteristics including age, sex, and household size.
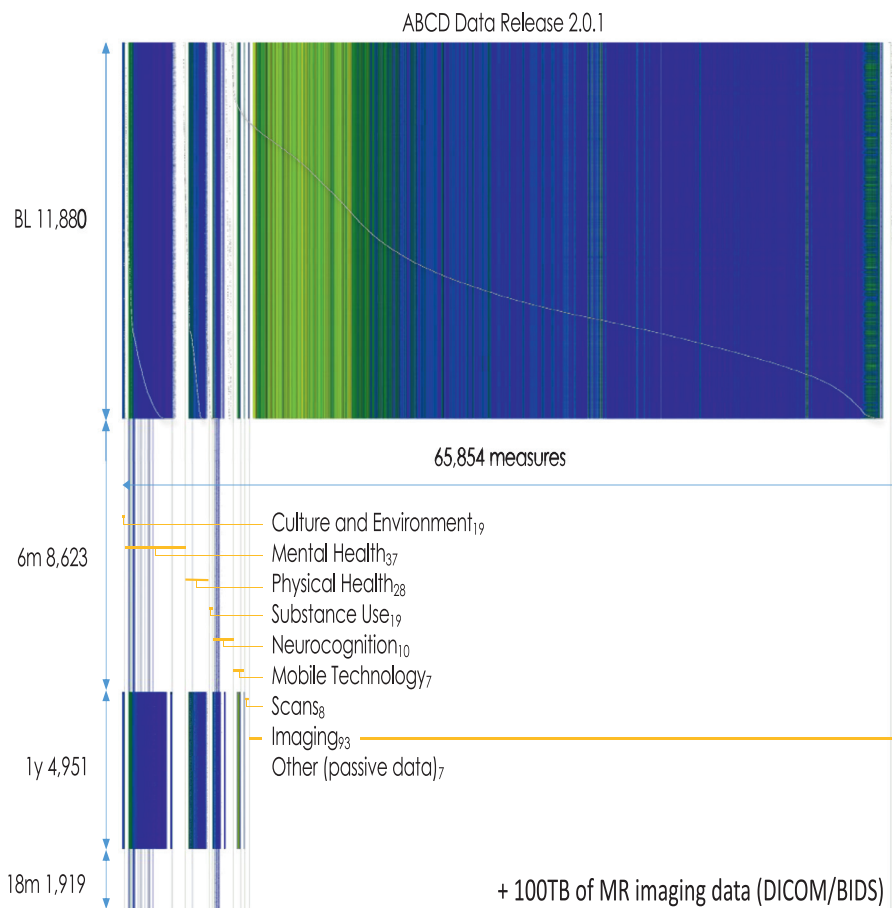
The general concordance of the samples can be attributed in large part to three factors: 1) the inherent demographic diversity across the ABCD study sites; 2) stratification (by race/ethnicity) in the probability sampling of schools within sites; and 3) demographic controls employed in the recruitment by site teams. Likewise, the unweighted percentages of ABCD children for the most prevalent race/ethnicity categories are an approximate match to the ACS estimates for US children age 9 and 10. Collectively, children of Asian, American Indian/Alaska Native (AIAN) and Native Hawaiian/Pacific Islander (NHPI) ancestry are under-represented in the unweighted ABCD data (3.2%) compared with ACS national estimates (5.9%). This outcome, which primarily affects ABCD's sample of Asian children, may be due in part to differences in how the parent/caregiver of the child reports multiple race/ethnicity ancestry in ABCD and the ACS.

3. Population inferences

The ABCD recruitment effort worked very hard to maintain similarity of the ABCD sample and the US population with respect to sex and race/ethnicity of the children in the study. The predominantly probability sampling methodology for recruiting children within each study site was intended to randomize over confounding factors that were not explicitly controlled (or subsequently reflected in the population weighting). Nevertheless, school consent and parental consent were strong forces that certainly may have altered the effectiveness of the randomization over these uncontrolled confounders.

3.1. Population weighting

The purpose of population weighting is to control for specific sources of selection bias and restore unbiasedness to descriptive and analytical estimates of the population characteristics and relationships (Heeringa et al., 2017). Briefly, construction of the population weights required identification of a key set of demographic and socioeconomic variables for the children and their households that are measured in both the ABCD Study and in the ACS household interviews. For the ABCD eligible children, the common variables include 1) age; 2) sex; and 3) race/ethnicity. For the child's household, additional variables include: 4) family income; 5) family type (married parents, single parent); 6) household size 7) parents' work force status (family type by parent employment status); 8) Census Region. A multiple logistic regression model using these variables was then fit to the concatenated ACS and ABCD data to predict study membership. The construction of the population weights for the ABCD Study is described in detail in Heeringa and Berglund (2020) (Heeringa and Berglund, 2020). R scripts for computing the ABCD population weights and for applying them in analyses are available at https://github.com/ABCD-STUDY/ abcd_acs_raked_propensity. The population weights are available in the NDA data releases 2.0.1 and 3.0.

3.2. Recommendations

Heeringa and Berglund (2020) (Heeringa and Berglund, 2020) present regression analyses with and without using the population weights. Although it is important not to over-generalize from a small set of comparisons to all possible analyses of the ABCD data, the results described therein lead to recommendations for researchers who are analyzing the ABCD baseline data. First, unweighted analysis may result in biased estimates of descriptive population statistics. The potential for bias in unweighted estimates from the ABCD data is strongest when the variable of interest is highly correlated with socioeconomic variables including family income, family type and parental work force

Release Year	Baseline	6 month	1 year	18 month	2 year	36 month	3 year	48 month	4 year	60 month	5 year	72 month	6 year	84 month	7 year	96 month	8 year	108 month	9 year	120 month	10 year	132 month	11 year	144 month	12 year
1	4,951	0																							
2	11,880	8,623	4,951	1,919	0																				
3	11,880	11,880	11,880	8,905	5,937	2,968	0																		
4	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0																
5	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0														
6	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0												
7	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0										
8	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0								
9	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0						
10	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0				
11	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0		
12	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937	2,968	0
13	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	8,905	5,937
14	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880	11,880
Collection year	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12

Yearly (rolling) release schedule

Baseline   Year 2   Year 4   Year 6   Year 8   Year 10

Spreadsheet data (year 12)
21 visits, 11,880 participants,
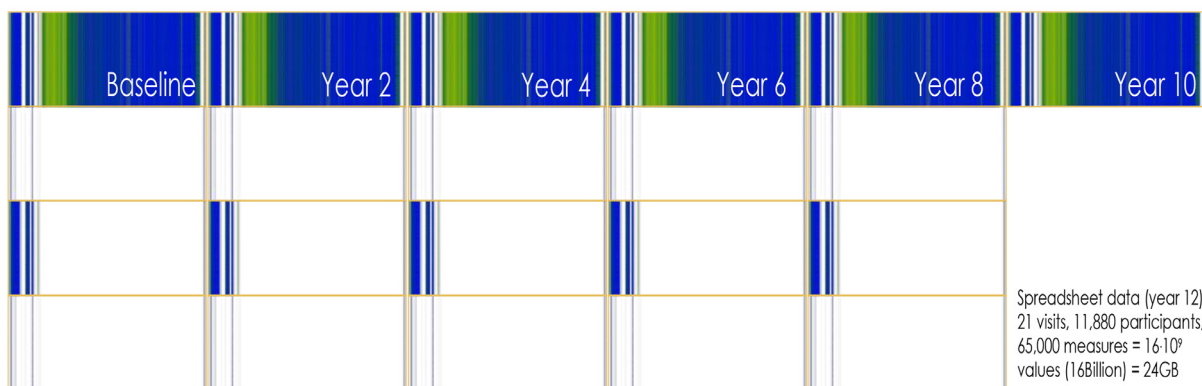65,000 measures = $16 \cdot 10^{9}$
values (16Billion) = 24GB

Fig. 2. ABCD Data Collection and NDA Release Schedule.

participation. Second, for regression models of the ABCD baseline data, an unweighted analysis using mixed-effects models (e.g., site, family, individual) is the preferred choice. Presently, there is no empirical evidence from comparative analyses that methods for multi-level weighting (Rabe-Hesketh and Skrondal, 2006) will improve the accuracy or precision of the model fit, although additional research on this topic is ongoing.

3.3. Example: Application to baseline brain volumes

As a demonstration of the implications of the weighting strategy employed in the ABCD Study, weighted and unweighted means and standard errors for ABCD baseline brain morphometry - volumes of cortical Desikan parcels (Desikan et al., 2006) - are presented in Table 2. Missing observations were first imputed using the R package **mice** (van Buuren and Groothuis-Oudshoorn, 2011) before applying weights to the completed sample. Differences between unweighted and weighted means are quite small in the baseline sample in this case. As longitudinal MRI data become available in ABCD (starting with the second post-baseline annual follow-up visit), population-valid mean trajectories of brain-related outcomes will also be computable using a similar population weighting scheme, also allowing for characterization of variation of trajectories from the population mean.

4. Hypothesis testing and power

Developing an operational approach to evaluate the meaningfulness of research findings has been a subject of consistent debate throughout

the history of statistics (Stigler, 1986). Even with the continued efforts to synthesize systems of statistical inference (Efron and Hastie, 2016), the resolution of this issue is unlikely to occur any time soon. Most neuroscientists continue to work within the context of the classical frequentist null-hypothesis significance testing (NHST) paradigm (Efron, 1998; Lehmann, 1993), although non-frequentist approaches (e.g. Bayesian, machine learning prediction (Efron, 2013; Efron, 2020)) are increasingly common and may be more appropriate for large datasets like the ABCD Study.

Despite growing enthusiasm for these alternatives, p-values continue to be important data points in the presentation of results in the behavioral and neurosciences. The NHST p-value "…is the probability under a specified statistical model that a statistical summary of the data…would be equal to or more extreme than its observed value" (Wasserstein and Lazar, 2016). The utility of NHST and the arbitrariness of the 0.05 significance threshold has been debated extensively (Gelman, 2018; Wasserstein and Lazar, 2016; Nickerson, 2000; Harlow et al., 2013). While we will not relitigate these issues here, we will attempt to address how best to present statistical evidence that leverages the ABCD Study's large sample size (affecting statistical power), population sampling frame, and rich longitudinal assessment protocol to enable meaningful and valid insights into child and adolescent neurodevelopment.

4.1. Power

Statistical power in the NHST framework is defined as the probability of rejecting a false null hypothesis. Power is determined by three

Table 1
ABCD Baseline and ACS 2011–2015 Demographic Characteristics.

Characteristic	Category	ABCD (n = 11,879)	ACS 2011–2015	
		%	N	%
Population	Total	100	8211,605	100
Age	9	52.3	4074,807	49.6
	10	47.8	4136,798	50.4
Sex	Male	52.2	4205,925	51.2
	Female	47.8	4005,860	48.8
Race/Ethnicity	NH White	52.2	4305,552	52.4
	NH Black	15.1	1101,297	13.4
	Hispanic	20.4	1973,827	24.0
	Asian, AIAN, NHPI	3.2	487,673	5.9
	Multiple	9.2	343,256	4.2
Family Income	<$25k	16.1	1762,415	21.5
	$25k - $49k	15.1	1784,747	21.7
	$50k - $74k	14.0	1397,641	17.0
	$75k - $99k	14.1	1023,127	12.5
	$100k - $199k	29.5	1685,036	20.5
	$200k +	11.2	558,639	6.8
Family Type	Married Parents	73.4	5426,131	66.1
	Other Family Type	26.6	2785,474	33.9
Parent Employment	Married, 2 in LF	50.2	3353,572	40.8
	Married, 1 in LF	21.9	1949,288	23.7
	Married, 0 in LF	1.3	156,807	1.9
	Single, in LF	21.1	2174,365	26.5
	Single, Not in LF	5.4	577,573	7.0
Region	Northeast	16.9	1336,183	16.3
	Midwest	20.4	1775,723	21.6
	South	28.3	3117,158	38.0
	West	34.4	1982,541	24.1
Household Size	2 to 3	17.3	1522,216	18.5
	4	33.5	2751,942	33.5
	5	24.9	2085,666	25.4
	6	14.0	1025,285	12.5
	7+	10.3	826,496	10.1

LF = labor force.
ACS = American Community Survey.

factors: 1) the significance level α; 2) the magnitude of the population parameter; and 3) the accuracy (precision and bias) of the model estimates. Increasing power while maintaining a specified Type I error rate depends largely on obtaining more precise association parameter estimates from improved study designs, more efficient statistical methods, and, importantly, increasing sample size (Rothman et al., 2008; Button et al., 2013; Hong and Park, 2012).

The ABCD Study has a large sample compared to typical neurodevelopmental studies, so much so that one might expect even very small associations to be statistically significant. In our experience, not all associations in the ABCD Study are guaranteed to have small p-values. For example, a recent study attempting to replicate the often-cited bilingual executive function advantage failed to find evidence for the advantage in the first data release (NDA 1.0) of the ABCD Study ($n = 4524$) (Dick et al., 2019).

Nevertheless, even very small associations are well-powered in the ABCD Study. Fig. 3 displays power curves as a function of sample size for different values of absolute Pearson correlations $|r|$. The dashed line in Fig. 3 indicates the full ABCD baseline sample size of $n = 11,880$. As can be seen, Pearson correlations $|r|=0.04$ and above have power > 0.99 at $\alpha = 0.05$. Simply rejecting a null hypothesis without reporting on other aspects of the study design and statistical analyses (including discussion of plausible alternative explanatory models and threats to validity), as well as the observed magnitude of associations, is uninformative, perhaps particularly so in the context of very well-powered studies (Abadie, 2020).

5. Effect sizes

Because p-values may be less informative in the context of very well-powered studies like ABCD, effect sizes become important data points

in determining the importance of the findings. Effect sizes quantify relationships between two or more variables, e.g., correlation coefficients, proportion of variance explained (R^2), Cohen's d, relative risk, number needed to treat, and so forth (Cohen, 1988; Kraemer, 1992; Rosenthal et al., 2000), with one variable often thought of as independent (exposure) and the other dependent (outcome) (Rothman et al., 2008). Effect sizes are independent of sample size, e.g., t-tests and p-values are not effect sizes; however, the precision of effect size estimators depend on sample size as described earlier. Consensus best practice recommendations are that effect size point estimates be accompanied by intervals to illustrate the precision of the estimate and the consequent range of plausible values indicated by the data (Wasserstein and Lazar, 2016). Table 3 presents a number of commonly used effect size metrics (Kirk, 1996; Fidler et al., 2004). We wish to avoid being overly prescriptive for which of these effect sizes to employ in ABCD applications, as researchers should think carefully about the intended use of their analyses and pick an effect size metric that addresses their particular research question.

5.1. Small effects

As much as the choice of which effect size statistic to report is driven by context, the interpretation of the practical utility of the observed effect size is even more so. While small p-values do not imply that reported effects are inherently substantive, "small" effect sizes might have practical or even clinical significance in the right context (Rosenthal et al., 2000).

As described in SM Section S.2, known problems of publication bias and incentives for researchers to find significant associations (Button et al., 2013; Simonsohn et al., 2014) combined with the predominantly small sample sizes of most prior neurodevelopmental stud-

Table 2
Unweighted and Weighted Means of Desikan Cortical Volumes.

	Mean	SE	Weighted Mean	SE
bankssts	3238.48	473.95	3227.7	472.83
caudalanteriorcingulate	2571.23	476.91	2559.34	478.06
caudalmiddlefrontal	8326.7	1408.47	8277.25	1398.77
cuneus	3645.25	582.41	3626.44	582.07
entorhinal	1843.15	339.44	1835.95	339.1
fusiform	12,050.11	1552.79	12,009.48	1558.06
inferiorparietal	18,387.31	2432.67	18,325.23	2428.86
inferiortemporal	13,182.85	1879.13	13,133.08	1870.21
isthmuscingulate	3252.16	534.48	3239.51	538.27
lateraloccipital	13,334.05	1870.71	13,283.9	1848.41
lateralorbitofrontal	9295.28	1036.65	9258.68	1035.6
lingual	8031.18	1132.35	7998.54	1132.13
medialorbitofrontal	5976.38	731.09	5954.65	725.41
middletemporal	14,275.5	1796.11	14,230.8	1786.83
parahippocampal	2586.48	378.94	2576.7	378.86
paracentral	4674.33	672.68	4660.61	674.3
parsopercularis	5701.08	849.03	5683.61	846.91
parsorbitalis	3097.73	371.12	3084.29	371.66
parstriangularis	5178.54	733.71	5159.42	732.41
pericalcarine	2505.86	425.52	2489.51	424.71
postcentral	11,822.49	1599.97	11,788.14	1593.43
posteriorcingulate	4196.07	603.72	4181.46	606.51
precentral	15,990.94	1796.68	15,929.85	1791.05
precuneus	12,865.56	1618.69	12,819.36	1616.69
rostralanteriorcingulate	2963.47	479.55	2949.78	479.97
rostralmiddlefrontal	21,292.13	2684.14	21,165.5	2669.35
superiorfrontal	28,758	3204.7	28,616.28	3197.22
superiorparietal	17,020.9	2172.8	16,961.33	2161.06
superiortemporal	14,575.38	1645.94	14,519.78	1652.24
supramarginal	13,827.92	1891.34	13,772.95	1894.8
frontalpole	1153.78	185.07	1150.68	186.2
temporalpole	2478.08	309.09	2472.2	308.04
transversetemporal	1339.14	216.87	1333.57	217.62
insula	7586.56	857.66	7556.2	856.7
total	297,024.1	28,733.9	295,831.8	28,686.9

Power as a Function of Sample Size for Pearson Correlation (r)

Fig. 3. Power vs. Sample Size for Pearson |r|.
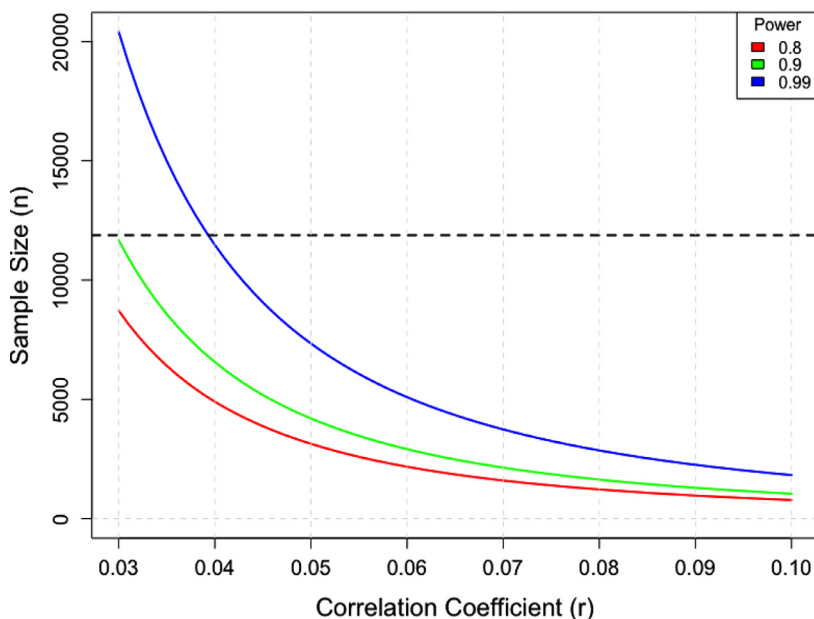
ies lead us to expect that true brain-behavior effect sizes are smaller than have been described in the past (Paulus and Thompson, 2019; Kendler, 2019). Indeed, Ioannidis (2005) (Ioannidis, 2005) has argued that most claimed research findings in the scientific literature are actually false. Although details of the concerns are disputed (Ashton, 2018), some analyses of existing literature provide support for the possibility (Bakker et al., 2012). It is possible, then, that many published neurodevelopmental associations represent severely inflated effect sizes (Button et al., 2013; Ioannidis, 2008) and may be severely attenuated in investigations of ABCD data.

Table 3
Measures of Effect Size Relevant for ABCD.

Measures of Strength of Association
r, r_{pb}, r^2, R, R^2, adjusted R^2, ϕ, η, η^2
Cohen's f^2
Cramér's V
Fisher's Z
Measures of Strength of Association Relevant for Multiple Regression
Standardized regression slope or path coefficient β
Semi-partial correlation $r_{y(x,z)}$
Measures of Effect Size
Cohen's d, f, g, h, q, w
Glass' g'
Hedges' g
Other Measures
Odds ratio (ω^2)
Relative risk

It is also possible that actual (causal) associations found in nature are in reality small for many outcomes. There is already strong evidence for this possibility: Myer and colleagues (2001) (Meyer et al., 2001) reviewed 125 meta-analyses in psychology and psychiatry and found that most relationships between clinically important variables are in the $r = 0.15$ to 0.3 range, with many clinically important effects even smaller. Miller et al. (2016) (Miller et al., 2016) analyzed associations between multimodal imaging and health-related outcomes in the UK-Biobank data. Even the most significant of these explained only around 1% of the variance in the outcomes.

5.2. Pre-Registration

While not of course completely immune to these problems (especially in subgroup and/or high-dimensional analyses), because its large sample size reduces random fluctuations in effect size estimates that occur within small n studies, the ABCD Study is much more resistant than is typical. However, with the large number of reseachers analyzing the data, high-dimensional space of covariates and outcomes and an essentially infinite number of possible modeling strategies, p-hacking and exploitation of random chance remains a possible source of irreproducible results. Pre-registration may mitigate exposure to some of these sources of irreproducibility. Indeed, a recent meta-analysis (Schäfer and Schwarz, 2019) found that effects from publications without pre-registration (median $r = 0.36$) skewed larger than effects from publications with pre-registration (median $r = 0.16$), suggesting that pre-registration is a practical step toward reporting research results that reflect the actual effects under investigation.

For ABCD Study data, we recommend that researchers consider hypothesis pre-registration (e.g., using the Open Science Foundation framework: https://osf.io/prereg/) and using a registered reports option for publishing results (Chambers et al., 2015). A template for hypothesis pre-registration for the ABCD Study data can be found in the NDA-hosted ABCD Data Exploration and Analysis Portal (ABCD DEAP, https://deap.nimhda.org/index.php), which is freely accessible to all users with a valid NDA ABCD user ID and password. Over 200 peer-review journals now offer registered reports as a publication format; two of these (*Cerebral Cortex* and *Developmental Cognitive Neuroscience*) have created registered reports options specifically geared for publishing results from the ABCD Study. Recommended best practices for promoting reproducible science are given in Section SM.5 and for statistical analyses and reporting of results using the ABCD Study data in Section SM.6. In the next section, we provide a brief example to illustrate the issues we have just discussed as they relate to ABCD.

5.3. Example: Effect size estimates

In examining the ABCD data, we advocate for a focus on effect sizes over p-values, but this is not as simple as it appears, and researchers of-

ten require some guidance on how to choose and interpret effect sizes. Here, we illustrate how the choice of effect size, and the interpretation of its substantive effect, must be made in the context of the research question. For example, Cohen's d and related metrics (see Table 3) assess the magnitude of mean differences between two conditions or groups. But what is not often appreciated is that Cohen's d is insensitive to the proportion of subjects in each group (McGrath and Meyer, 2006). Conversely, base-rate-sensitive effect size metrics take into account the difficulty of differentiating phenomena in rare events. If the goal is to assess the impact of an exposure on a population, it is arguable that researchers should opt for an effect size metric that takes the sample base rate into account. For example, the point-biserial correlation r_{bs} (McGrath and Meyer, 2006) (Table 3) is a similar metric that, unlike d, is sensitive to variation in sample base rates.

To illustrate this, we used Cohen's d and point-biserial r_{bs} to estimate the effect size of a dichotomous "exposure" index: very obese (here defined as a body mass index (BMI) ≥ 30) and a continuous brain "outcome": restriction spectrum imaging component (N0), a measure sometimes related to cellularity, in the Nucleus Accumbens (NAcc). Recent work has highlighted a potential role of neuroinflammation in the NAcc in animal models of diet-induced obesity (Décarie-Spain et al., 2018). We included baseline data from subjects without missing BMI and NAcc N0 data, also excluding 5 subjects with NAcc N0 values < 0 (leaving $n = 10,659$ subjects, of which 184 subjects had BMI ≥ 30, or 1.7%). As can be seen in Fig. 4 (upper panels), NAcc N0 values are heavy tailed. We thus use a bootstrap hypothesis testing procedure to obtain quantiles of d and r_{bs} (Martin, 2007). To account for nesting of subjects within families, at each iteration of the bootstrap one member of each family was first selected at random, and these subjects (along with all singletons) were sampled with replacement 10,000 times. Fig. 4 (lower panels) presents the bootstrap p-value plots for different null hypotheses (Rothman et al., 2008). The bootstrap median $d = 0.801$ (95% CI: [0.588, 0.907]) and median $r_{bs} = 0.106$ [0.081, 0.127]. Thus, while in terms of d the effect might be considered "large", r_{bs} corresponds to a variance explained of roughly 1% and hence would be considered "small" by many researchers.

So, what effect size should the researcher report, and which should be emphasized in the interpretation? Our general guidance would be to carefully consider the answer in the context of the research question. Perhaps both could be reported, but if the public health impact of an intervention is considered the r_{bs} might be more strongly focused on in the discussion of results.

Finally, caution is warranted in interpreting these results as "effect sizes," as the causal relationship could be from obesity to NAcc N0, from NAcc N0 to obesity, bidirectional, or even non-existent (i.e., due to confounding). We do not adjust for potential confounding factors or their proxies in this example. In light of this, it would be more appropriate to call d and r_{bs} as computed here "association sizes". We examine the question of direction of causality using the twin data (Heath et al., 1993) in SM Section S.3.

6. Control of confounding variables

An important challenge to the internal validity of effect estimates from the ABCD Study (and from any observational study) is the likely presence of confounding variables for observed associations. Necessary but not sufficient conditions for a variable to confound an observed association between an independent variable (IV) and a dependent variable (DV) are that the factor is associated with both the exposure and the outcome in the population, but not causally affected by either (VanderWeele and Shpitser, 2013) (if a variable is causally downstream of the IV or the DV or both, it may be a collider or a mediator (Rothman et al., 2008)). Conditioning on confounders (or their proxies) in regression analyses will tend to reduce bias in effect size estimates, whereas conditioning on colliders or mediators (or their proxies) will tend to increase bias. To make matters more difficult, assessed variables
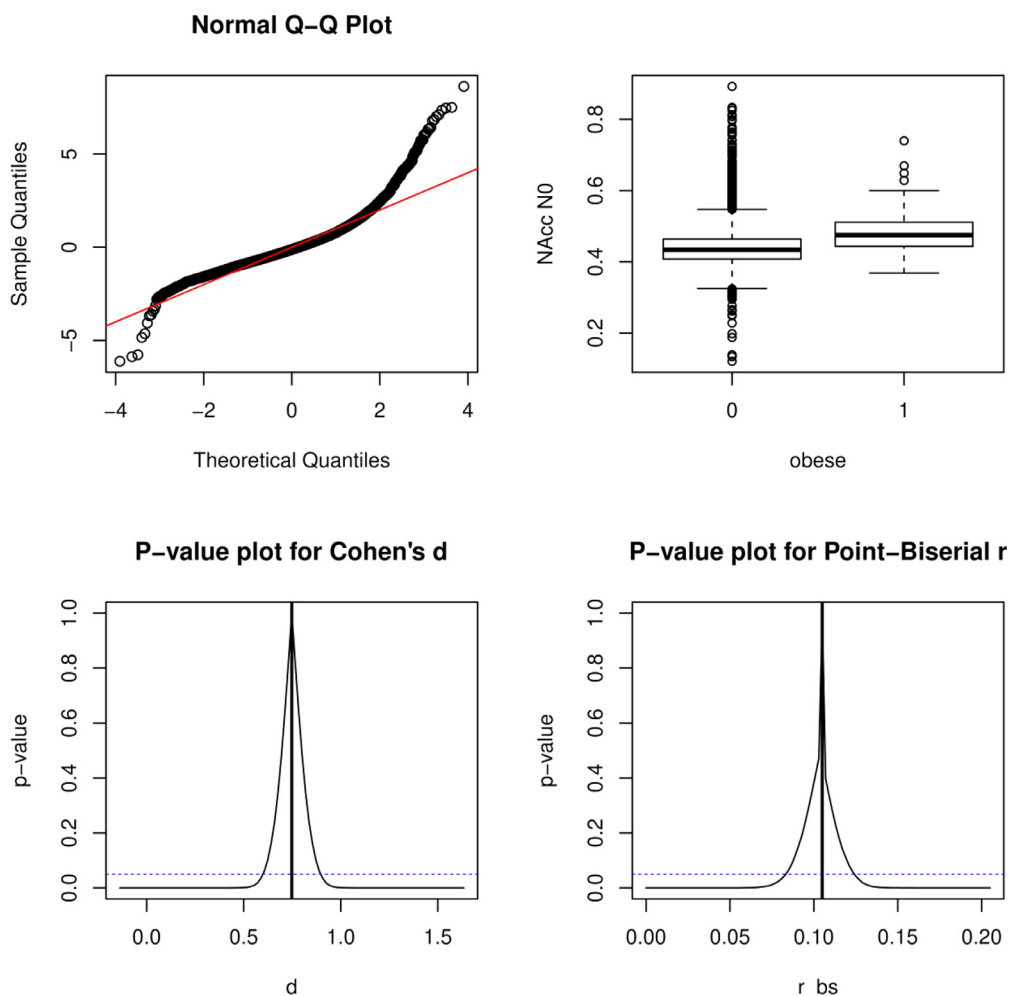
Normal Q–Q Plot

P–value plot for Cohen's d

P–value plot for Point–Biserial r

can be proxies for both confounding factors and mediators or colliders simultaneously, in which case it is not clear whether conditioning will improve or worsen bias in effect size estimates. We thus recommend that investigators using ABCD data think carefully about challenges to estimating effects of exposures and perform sensitivity analyses that examine the impact of including/excluding covariates on associations. In the next sections we discuss these topics more thoroughly in the context of conditioning on covariates in regression models.

6.1. Covariate adjustment

Although the inclusion of covariates in statistical models is a widespread practice, determining which covariates to include is necessarily complex and presents an analytical conundrum. The advantages and disadvantages of covariate inclusion in statistical models has been widely debated (Meehl, 1971; Schwarz, 1970) and reviewed elsewhere (Atinc et al., 2012; Becker et al., 2016; Spector and Brannick, 2011), so we focus our discussion on the practical implications of covariate adjustment in the ABCD Study.

Datasets with a rich set of demographic and other variables lend themselves to the inclusion of any number of covariates. In many respects, this can be seen as a strength of the ABCD Study, but this can also complicate the interpretation of findings when research groups adopt different strategies for what covariates to include in their models. For instance, a recent comprehensive review of neuroimaging studies (Hyatt et al., 2020) found that the number of covariates used in models ranged from 0 to 14, with 37 different sets of covariates across the 68 models reviewed. This review showed that brain-behavior associations

varied substantially as a function of which covariates were included in models: some sets of covariates influenced observed associations only a little, whereas others resulted in dramatically different patterns of results compared to models with no covariates. Which variables are appropriately included as confounders in any given analysis depends on the research question, highlighting the need for thoughtful use of covariates.

Covariates are often used in an attempt to yield more "accurate," or "purified" (Spector and Brannick, 2011) estimates of the relationships among the IVs and DV, thereby revealing their "true" associations (Atinc et al., 2012) (i.e., to eliminate the impact of confounding on observed associations (Rothman et al., 2008)). Under this assumption, the inclusion of covariates implicitly assumes that they are somehow influencing the variables of interest, either contaminating the relationship between the IV and DV or the measurement of the variables of interest. Thus, not controlling for covariates presumably distorts observed associations among the IVs and DV (Meehl, 1971; Spector and Brannick, 2011). Note that we use "somehow" to emphasize frequent researcher agnosticism regarding the specific role of the covariates included in the model. Because statistical control carries with it major assumptions about the relationships among the observed variables and latent constructs, some of which are generally unspecified and others of which are potentially unknowable, conclusions drawn from models that mis-specify the role of the covariate will be incorrect.

When covariates are thought to influence the observed variables of interest but not the latent construct, this is thought of as measurement contamination (Fig. 5A). Measurement contamination ostensibly occurs when a covariate influences the observed variables (x and y in Fig. 5A).
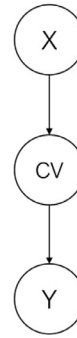
(A) Measurement contamination
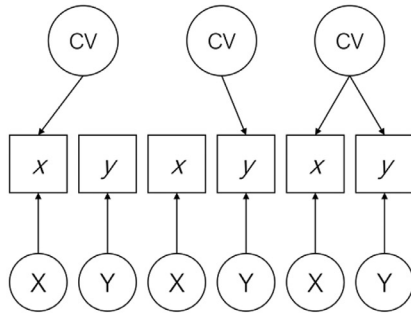
(B) Spuriousness
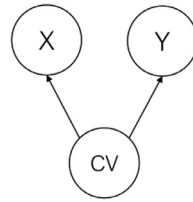
(C) Mediation

Fig. 5. Models for Measurement Contamination, Spuriousness, and Mediation.

Importantly, a major assumption surrounding the presumption of measurement contamination is that the covariate does not affect the underlying constructs (X and Y in Fig. 5A), only their measures. Removing the influence of covariates by controlling for them presumes that absent such control, the association between the IVs and DV is artefactual.

There are also a number of ways in which covariates are thought to influence the latent constructs and not just the measurement of them (see Meehl (1971) (Meehl, 1971) for a thorough discussion). Two such models are spuriousness (Fig. 5B) and mediation (Fig. 5C). Under a spuriousness (confounding) model, the IV (X) and DV (Y) are not directly causally associated but are both caused by the covariate. Therefore, any observed association between the IV and DV is spurious given that it is caused by the covariate. Under a mediation model, the IV (X) and DV (Y) are statistically associated only through the covariate. Spuriousness and mediation models are generally statistically indistinguishable (though temporal ordering can sometimes assist in appropriate intepretations), and under both models, controlling for the covariate results in a reduced association between the IV and DV. In either case, including covariates can effectively remove effects of interest from the model. At best, this practice obscures rather than purifies relationships among our variables of interest. At worst, this practice can render incorrect interpretations of the true effect. Rather than suggesting that covariates should be avoided altogether, we view them as having an important role in testing competing hypotheses.

Thorough treatments of covariate use in statistical modeling are given by others (Atinc et al., 2012; Becker et al., 2016; Spector and Brannick, 2011). In the next section we review steps in reasoning about which covariates to include and how to think about resulting associations.

6.1.1. Covariate adjustment: researcher considerations

What is the role of the covariate? What is the theoretical model? Could the exclusion and inclusion of the covariate inform the theoretical model? Addressing these questions through the practice of simply explicitly specifying the role of the covariate in the model, and even more specifically its hypothesized role in the IV-DV associations, helps avoid including covariates in the model when doing so is poorly justified. Moreover, it encourages thoughtful hypothesis testing. Ideally, explicit justification of the inclusion of each covariate in the model should be included in the reporting of our results. Better yet, as opposed to treating control variables as nuisance variables, a more ideal model would include covariates in hypotheses (Becker et al., 2016). We also encourage considering the extent to which the exclusion and inclusion of the covariate could inform the theoretical model. In an explanatory framework, all covariates should be specified *a priori*. In a predictive framework, one can conduct nested cross-validations and model comparisons to find the most robust model with procedurally-selected covariates.

How do my models differ with and without covariates? We recommend running models with and without covariates and comparing their results. This practice encourages researchers to better consider the effect

of covariates on observed associations. At the same time, engaging in multiple testing can increase Type I error rates. Regarding our suggestion, we encourage a shift away from comparing models on the basis of p-values and instead encourage researchers to compare effect sizes of the predictor of interest in models with and without the covariates. Confidence intervals are critical to compare across models, as the range of plausible effects is more important than the point differences in effect size estimates. The focus on effect sizes as opposed to statistical significance is important given that including many covariates in the statistical model reduces degrees of freedom, in turn increasing standard errors and decreasing statistical power for any given IV. Covariates may be correlated with one another as well, reducing precision and producing large differences in p-values when some variables are included or omitted from a model.

If the effect sizes do not differ as a function of the inclusion of the covariate (e.g., their confidence intervals substantially overlap), one might consider dropping it from the model, but noting this information somewhere in the text. Becker (2005) (Becker, 2005) offers more suggestions regarding what to do when results from models with and without covariates differ (see also Becker et al. (2016) (Becker et al., 2016)). Additionally, should one choose to adopt models with covariates included, we recommend placing analyses from models without covariates in an appendix or in the supplemental materials; such a practice will aid in comparison of results across studies, particularly across studies with different sets of covariates in the models.

Causal effects from observational data It is worth formalizing this discussion for situations when there is interest in estimating causal effects: the comparison of potential outcomes, e.g., comparing outcomes for children in ABCD as if all of their parents had alcohol problems, vs. none of their parents having alcohol problems. Two methods that are particularly relevant for estimating causal effects in cohort studies such as ABCD are instrumental variables analyses and propensity score methods. Instrumental variables analyses rely on finding some "instrument" that is plausibly randomly assigned (conditional on covariates), affects the exposure of interest, and is not directly related to outcomes (Angrist et al., 1996; Hernán and Robins, 2006).

Here we will focus, though, on propensity score methods as a fairly general purpose tool for estimating causal effects. In general, interpreting a difference in outcomes between exposure groups as a causal effect requires two things: 1) "overlap" (individuals in the two exposure groups are similar to one another on the confounders), and 2) "unconfounded treatment assignment"; that there are no unobserved differences between exposure groups once the groups are equated on the observed characteristics. Propensity score methods (Stuart, 2010) can help assess whether overlap exists, and equate the exposure groups using matching, weighting, or subclassification. Covariates should thus be selected in order to satisfy unconfounded treatment assignment, and as noted above, factors that are "post-treatment" (and thus potentially mediators) should not be included. A benefit of the ABCD Study design is that longitudinal data is available, to measure confounders be-

fore exposure and exposure before outcomes, and the large set of potential confounders observed and available to be adjusted for. Sensitivity analyses also exist to assess robustness of effect estimates to a potential unobserved confounder (e.g., (Cinelli et al., 2020; Liu et al., 2013; VanderWeele and Ding, 2017)).

In SM Section S.4, we give a worked example of a sensitivity analysis for the potential impact of omitting unmeasured confounders using ABCD data on breastfeeding and neurocognition. Finally, methods should be used that account for the probability sample nature of the ABCD cohort, in order to ensure effects are being estimated for the population of interest (Lenis et al., 2019; Ridgeway et al., 2015).

6.1.2. Example: Covariate adjustment

Here, we provide a worked example focusing on the associations between parental history of alcohol problems and child psychopathology, an important substantive question that has received attention in the literature (Hesselbrock and Hesselbrock, 1992). The ABCD Study contains a rich assessment of family history of psychiatric problems (e.g., alcohol problems, drug problems, trouble with the law, depression, nerves, visions, suicide) and child psychopathology, including child- and parent-reported dimensional and diagnostic assessments. We examined the relation between parental history of alcohol problems (four levels: neither parent with alcohol problems, father only, mother only, both parents) and child externalizing assessed with the parent-reported Child Behavior Checklist (CBCL). Based on the earlier-described considerations, we delineated several tiers of covariates to include in the models in sequence (or in a stepwise fashion). The first tier included "essential" covariates that the researcher views as required to include in the models, the second tier included "non-essential" covariates, and the third tier included "substantive" covariates that can inform the robustness of the model, or more generally inform the theoretical model.

Our first tier includes age, sex at birth, and a composite of maternal alcohol consumption while pregnant. The inclusion of this latter covariate is deemed as essential to rule out the possibility that any associations between parental history of alcohol problems and child psychopathology was not due to prenatal alcohol exposure. In this context, maternal alcohol consumption was considered a construct confound. The second-tier covariates included race/ethnicity, household income, parental education, and parental marital status. In the context of this research question, these covariates might be deemed "non-essential" for three reasons. First, the researcher may not have any clear hypotheses surrounding the role of these covariates in the IV-DV associations. Second, the researcher may not think that there are important group differences in the second-tier covariates that are worth exploring and reporting. Third, the researcher might expect that some of the "non-essential" covariates may be causally related to the IVs and DV or may share common causes with them (e.g., they may be proxies for both confounders and mediators or colliders simultaneously). We did not have specific hypotheses regarding race/ethnicity differences in these associations, but exploratory analyses may be informative. At the same time, race/ethnicity may be strongly associated with other covariates (e.g., socioecomomic status, adversity), and so researchers must take care when interpreting the impact of its inclusion in the model.

Other "non-essential" covariates (e.g., household income, parental education, and parental marital status) may be either causally related to the IVs or DV or may share a common cause. For instance, parental externalizing – which likely overlaps with parental history of alcohol problems – are associated with both increased likelihood of divorce and child externalizing, but the two are not causally related (Lahey et al., 1988; Salvatore et al., 2017). Thus, demographics may, at least in part, proxy our variables of interest. Moreover, parental history of alcohol problems may proxy the broader construct of externalizing psychopathology. Controlling for indicators that share a common cause with our IVs and DVs partials out an important, etiologically relevant part of the phenotype, which can obscure true IV-DV associations.Based on this information, one might decide to report models with and without these covariates

and consider the extent to which differences in these sets of models inform a particular theoretical model.

There was a significant linear association between parental history of alcohol problems with tier 1 covariates included, and there is no major difference between the models with and without tier 2 covariates (Fig. 6A). Because we deemed tier 2 covariates as "nonessential," we elected to move forward only with tier 1 covariates.

Finally, a third tier of covariates may be used to test the robustness of the associations between parental history of alcohol problems and child psychopathology. Here, we see that other forms of parental history of psychiatric problems, particularly externalizing (i.e., parental history of drugs, trouble with the law) display similar, if not more robust associations, with CBCL Externalizing (Fig. 6B). Including other forms of parental externalizing (e.g., drug use, trouble with the law), may inform the extent to which the associations between parental history of alcohol problems and child psychopathology are more general to parental history of other externalizing (Kendler et al., 2011)). Indeed, the associations between parental history of alcohol problems and CBCL Externalizing became attenuated when parental history of drug problems and trouble with the law were included in the model (Fig. 6C), which suggests that the associations are general with respect to parental history of externalizing. In one further robustness check, we see that including parental history of internalizing problems (e.g., nerves, depression) slightly attenuates the associations between parental history of alcohol problems and CBCL Externalizing, though the effects of covarying parental history of externalizing were stronger (Fig. 6C).

Altogether, we learned from the tier 3 covariates that the associations between parental history of alcohol problems and CBCL Externalizing may be more general to history of externalizing, or even psychiatric problems more generally. These covariates were not treated as covariates *per se*, but as variables whose inclusion and exclusion informed the theoretical model.

In sum, we hope it is clear that determining which covariates should be included in our statistical models is complex and requires considerable thought. We caution against the over-inclusion of covariates in statistical models, and against the assumption that including covariates purifies the associations among our variables of interest; instead their inclusion can obscure rather than purify such associations (Schisterman et al., 2009).

7. Summary and conclusions

The sample size of the ABCD Study is large enough to reliably detect and estimate small effect size relationships among a multiplicity of genetic and environmental factors, potential biological mechanisms, and behavioral and health-related trajectories across the course of adolescence. Thus, the ABCD Study will be a crucial resource for avoiding Type I errors (false positive findings) when discovering novel relationships, as well as failures to replicate that result from the replication sample being too small to have sufficient power. Moreover, ABCD will allow for stronger interpretation of non-significant results as they will not be due to low power for all but the tiniest of effect sizes, or researchers may opt to take advantage of the high power to assess the absence of differences using other statistical procedures like equivalence tests (Lakens, 2017). Other studies in the field suffer from false positives that do not replicate, and overestimation of effect sizes in general, which typically arise from a research environment consisting of many small studies, p-hacking, and publication bias towards positive findings (Walum et al., 2016). ABCD will therefore help directly address the replication problems afflicting much of current neuroscience research (Button et al., 2013), and which would be bolstered by pre-registration steps that we outline above for ABCD data.

ABCD may also help researchers to address questions of "practical significance" for effects that may be small by traditional standards (e.g., explaining 1% of variation or less), but may be statistically significant
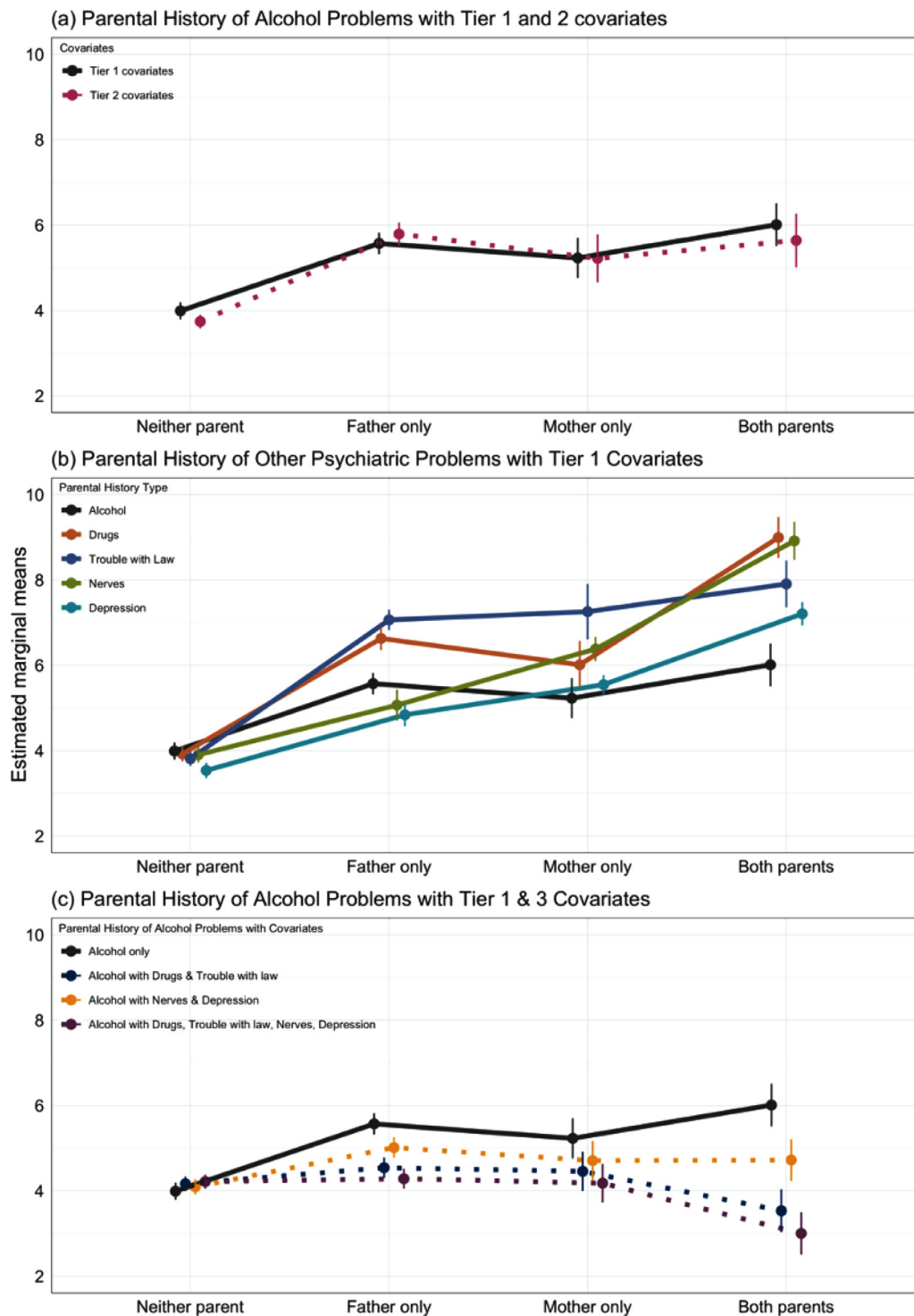
Fig. 6. The association between parental history of alcohol problems and CBCL Externalizing.

due to the large sample size of the ABCD Study. As we noted, we expect that ABCD report will predominantly report small effect sizes, simply reflecting the fact that many, if not most, real-world relationships are in fact small. But in this scenario, it would be a mistake to dismiss all small effect size relationships. Indeed, an ostensibly small effect size might still be of clinical or public health interest (Rosenthal et al., 2000) despite appearing "small" by traditional standards (McClelland and Judd, 1993; Wray et al., 2018). The effect may also be small due to imprecise measurement even if the underlying relationships are far from weak. Finally, even if the "noise-free" effects of individual factors are small, they may

cumulatively explain a sizeable proportion of the variation in neurodevelopmental trajectories a scenario which has recently played out in genome-wide association studies (GWAS) of complex traits (Boyle et al., 2017).

At the same time, it is important to interpret these effects in the context of potentially confounding covariates, and like the interpretation of the effects themselves, the choice of inclusion of covariates must be principled. Misspecification can lead to serious threats to internal validity of the conclusions. For both effects of primary interest and covariates, that the focus remains on effect sizes, rather than binary "yes or no"

assessments of whether data support or reject a particular hypothesis. For example, for the goal of obtaining personally relevant modifiable predictors of substance abuse or other clinical outcomes, prediction accuracy of 75% would correspond to a very-large effect size of around 1.4, accounting for about 30% of the variance. (However, for modifications of variables targeted at a population level or for policy interventions, a smaller effect size might still be important.) Thus, binary judgements on whether associations are "significant" can be fraught with error and give rise to misleading headlines (Goodman, 2008). Worse, Type I or Type II errors (declaring an effect to be significant when it is not real, or absent when it is, respectively) can mislead the field for long periods. Such results could delay the much needed progress in reducing the human and financial costs of mental health and other disorders. Thus, the careful consideration of the statistical and methodological factors we have outlined should be considered essential for the investigation of this prominent public dataset.

In summary, the ABCD Study is collecting longitudinal data on a rich variety of genetic and environmental data, biological samples, markers of brain development, substance use, and mental and physical health, enabling the construction of realistically complex etiological models incorporating factors from many domains simultaneously. While establishing reproducible relationships between pairs (or small collections of measures) in a limited set of domains will still be important, it will be crucial to develop more complex models from these building blocks to explain enough variation in outcomes to reach a more complete understanding or to obtain clinically-useful individual predictions. Multidimensional statistical models must then incorporate knowledge from a diverse array of domains (e.g., genetics and epigenetics, environmental factors, policy environment, ecological momentary assessment, school-based assessments, and so forth) with brain imaging and other biologically-based measures, behavior, psychopathology, and physical health, and do this in a longitudinal context. The sample size, population nature, duration of study, and, importantly, the richness of data collected in ABCD will be important for attaining this goal.

Data and Code Availability Statement

Data are publicly released on an annual basis through the NIMH Data Archive (NDA, https://nda.nih.gov/abcd). The ABCD Study data are openly available to qualified researchers for free. Access can be requested at https://nda.nih.gov/abcd/request-access. Code for the replication of analyses conducted in the manuscript can be retrieved at https://github.com/ABCD-STUDY/

Acknowledgments

ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from NIMH Data Archive Release 2.0.1 (DOI 10.15154/1506087). DOIs can be found at https://nda.nih.gov/abcd.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118262.

References

Abadie, A., 2020. Statistical nonsignificance in empirical economics. Am. Econ. Rev.: Insights 2, 193–208.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Am. Statist. Assoc. 91, 444–455.

Ashton, J.C., 2018. It has not been proven why or that most research findings are false. Med. Hypotheses 113, 27–29.

Atinc, G., Simmering, M.J., Kroll, M.J., 2012. Control variable use and reporting in macro and micro management research. Org. Res. Methods 15, 57–74.

Bachman, J.G., Johnston, L.D., O'Malley, P.M. & Schulenberg, J.E. The monitoring the future project after thirty-seven years: design and procedures. (2011).

Bagot, K., et al., 2018. Current, future and potential use of mobile and wearable technologies and social media data in the abcd study to increase understanding of contributors to child health. Dev. Cognit. Neurosci. 32, 121–129.

Bakker, M., Dijk, A.van, Wicherts, J.M., 2012. The rules of the game called psychological science. Perspect. Psychol. Sci. 7, 543–554.

Barch, D.M., et al., 2018. Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: rationale and description. Dev. Cognit. Neurosci. 32, 55–66.

Becker, T.E., 2005. Potential problems in the statistical control of variables in organizational research: a qualitative analysis with recommendations. Org. Res. Methods 8, 274–289.

Becker, T.E., et al., 2016. Statistical control in correlational studies: 10 essential recommendations for organizational researchers. J. Organ. Behav. 37, 157–167.

Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An expanded view of complex traits: from polygenic to omnigenic. Cell 169, 1177–1186.

Button, K.S., et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365.

Casey, B., et al., 2018. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. Dev. Cognit. Neurosci. 32, 43–54.

Chambers, C.D., Dienes, Z., McIntosh, R.D., Rotshtein, P., Willmes, K., 2015. Registered reports: realigning incentives in scientific publishing. Cortex 66, A1–A2.

Chantala, K., Tabor, J., 1999. National longitudinal study of adolescent health. Strategies to perform a design-based analysis using the add health data.

Cinelli, C., Ferwerda, J., Hazlett, C., 2020. Sensemakr: sensitivity analysis tools for ols in r and stata. Submitted to the J. Stat. Softw..

Cohen, J. Statistical power analysis. (1988).

Conway, K.P., Swendsen, J., Husky, M.M., He, J.-.P., Merikangas, K.R., 2016. Association of lifetime mental disorders and subsequent alcohol and illicit drug use: results from the national comorbidity survey–adolescent supplement. J. Am. Acad. Child Adolesc. Psychiatry 55, 280–288.

Décarie-Spain, L., et al., 2018. Nucleus accumbens inflammation mediates anxiodepressive behavior and compulsive sucrose seeking elicited by saturated dietary fat. Mol. Metabol. 10, 1–13.

Desikan, R.S., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31, 968–980.

Dick, A.S., et al., 2019. No evidence for a bilingual executive function advantage in the abcd study. Nature Human Behav. 3, 692–701.

Efron, B., 1998. RA fisher in the 21st century. Stat. Sci. 95–114.

Efron, B., 2013. Bayes' theorem in the 21st century. Science 340, 1177–1178.

Efron, B., 2020. Prediction, estimation, and attribution. J. Am. Statist. Assoc. 115, 636–655.

Efron, B., Hastie, T., 2016. Computer Age Statistical Inference. Cambridge University Press, p. 5.

Fidler, F., Thomason, N., Cumming, G., Finch, S., Leeman, J., 2004. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol. Sci. 15, 119–126.

Garavan, H., et al., 2018. Recruiting the abcd sample: design considerations and procedures. Dev. Cognit. Neurosci..

Gelman, A., 2018. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Pers. Soc. Psychol. Bull. 44, 16–23.

Goodman, S., 2008. A dirty dozen: twelve p-value misconceptions. Semin. Hematol. 45, 135–140 Elsevier.

Hagler, D.J., et al., 2018. Image processing and analysis methods for the adolescent brain cognitive development study. bioRxiv 457739.

Harlow, L.L., Mulaik, S.A., Steiger, J.H., 2013. What If There Were No Significance tests?. Psychology Press.

Heath, A.C., et al., 1993. Testing hypotheses about direction of causation using cross-sectional family data. Behav. Genet. 23, 29–50.

Heeringa, S.G., Berglund, P.A., 2020. A guide for population-based analysis of the adolescent brain cognitive development (abcd) study baseline data. BioRxiv.

Heeringa, S.G., West, B.T., Berglund, P.A., 2017. Applied Survey Data Analysis. Hall/CRC, Chapman.

Hernán, M.A., Robins, J.M., 2006. Instruments for causal inference: an epidemiologist's dream? Epidemiology 360–372.

Hesselbrock, M.N., Hesselbrock, V.M., 1992. Relationship of family history, antisocial personality disorder and personality traits in young men at risk for alcoholism. J. Stud. Alcohol 53, 619–625.

Hong, E.P., Park, J.W., 2012. Sample size and statistical power calculation in genetic association studies. Genom. Informatics 10, 117.

Hyatt, C.S., et al., 2020. The quandary of covarying: a brief review and empirical examination of covariate use in structural neuroimaging studies on psychological variables. Neuroimage 205, 116225.

Iacono, W.G., et al., 2017. The utility of twins in developmental clinical neuroscience research: how twins strengthen the abcd research design. Dev. Cognit. Neurosci..

Ingels, S., Abraham, S., Karr, R., Spenser, B., Frankel, M., 1990. National Education Longitudinal Survey of 1988. National Opinion Research Center, University of Chicago *Technical Report.*

Ioannidis, J.P., 2005. Why most published research findings are false. PLos med 2, e124.

Ioannidis, J.P., 2008. Why most discovered true associations are inflated. Epidemiology 640–648.

Kendler, K.S., et al., 2011. The structure of genetic and environmental risk factors for syndromal and subsyndromal common dsm-iv axis i and all axis ii disorders. Am. J. Psychiatry 168, 29–39.

Kendler, K.S., 2019. From many to one to many—The search for causes of psychiatric illness. JAMA Psychiatry 76, 1085–1091.

Kirk, R.E., 1996. Practical significance: a concept whose time has come. Educ. Psychol. Meas. 56, 746–759.

Kraemer, H.C., 1992. Reporting the size of effects in research studies to facilitate assessment of practical and clinical significance. Psychoneuroendocrinology 17, 527–536.

Lahey, B.B., et al., 1988. Conduct disorder: parsing the confounded relation to parental divorce and antisocial personality. J. Abnorm. Psychol. 97, 334.

Lakens, D., 2017. Equivalence tests: a practical primer for t tests, correlations, and meta–analyses. Social Psychol. Pers. Sci. 8, 355–362.

Lehmann, E.L., 1993. The fisher, neyman-pearson theories of testing hypotheses: one theory or two? J. Am. Statist. Assoc. 88, 1242–1249.

Lenis, D., Nguyen, T.Q., Dong, N., Stuart, E.A., 2019. It's all about balance: propensity score matching in the context of complex survey data. Biostatistics 20, 147–163.

Liu, W., Kuramoto, S.J., Stuart, E.A., 2013. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prev. Sci. 14, 570–580.

Loughnan, R., et al., 2020. Polygenic score of intelligence is more predictive of crystallized than fluid performance among children. bioRxiv 637512.

Luciana, M., et al., 2018. Adolescent neurocognitive development and impacts of substance use: overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery. Dev. Cognit. Neurosci..

Martin, M.A., 2007. Bootstrap hypothesis testing for some common statistical problems: a critical evaluation of size and power properties. Comput. Stat. Data Anal. 51, 6321–6342.

McClelland, G.H., Judd, C.M., 1993. Statistical difficulties of detecting interactions and moderator effects. Psychol. Bull. 114, 376.

McGrath, R.E., Meyer, G.J., 2006. When effect sizes disagree: the case of r and d. Psychol. Methods 11, 386.

Meehl, P.E. High school yearbooks: a reply to schwarz. (1971).

Meyer, G.J., et al., 2001. Psychological testing and psychological assessment: a review of evidence and issues. Am. Psychol. 56, 128.

Miller, K.L., et al., 2016. Multimodal population brain imaging in the uk biobank prospective epidemiological study. Nat. Neurosci. 19, 1523.

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol. Methods 5, 241.

Paulus, M.P., Thompson, W.K., 2019. The challenges and opportunities of small effects: the new normal in academic psychiatry. JAMA Psychiatry 76, 353–354.

Rabe-Hesketh, S., Skrondal, A., 2006. Multilevel modelling of complex survey data. J. R. Stat. Soc.: Series A (Statistics in Society) 169, 805–827.

Ridgeway, G., Kovalchik, S.A., Griffin, B.A., Kabeto, M.U., 2015. Propensity score analysis with survey weighted data. J. Causal Inference 3, 237–249.

Rosenthal, R., Rosnow, R.L., Rubin, D.B., 2000. Contrasts and Effect Sizes in Behavioral research: A correlational Approach. Cambridge University Press.

Rothman, K.J., Greenland, S., Lash, T.L., 2008. Modern Epidemiology. Lippincott Williams & Wilkins.

Salvatore, J.E., et al., 2017. Alcohol use disorder and divorce: evidence for a genetic correlation in a population-based swedish sample. Addiction 112, 586–593.

Schäfer, T., Schwarz, M.A., 2019. The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. Front. Psychol. 10, 813.

Schisterman, E.F., Cole, S.R., Platt, R.W., 2009. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 20, 488.

Schwarz, J.C., 1970. Comment on" high school yearbooks: a nonreactive measure of social isolation in graduates who later became schizophrenic. J. Abnorm. Psychol. 75, 317.

Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. P-curve: a key to the file-drawer. J. Exper. Psychol.: General 143, 534.

Spector, P.E., Brannick, M.T., 2011. Methodological urban legends: the misuse of statistical control variables. Org. Res. Methods 14, 287–305.

Stigler, S.M., 1986. The History of statistics: The measurement of Uncertainty Before 1900. Harvard University Press.

Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Stat. Sci.: Rev. J. Inst. Math. Stat. 25, 1.

Thompson, W.K., et al., 2019. The structure of cognition in 9 and 10 year-old children and associations with problem behaviors: findings from the abcd study's baseline neurocognitive battery. Dev. Cognit. Neurosci. 36, 100606.

Uban, K.A., et al., 2018. Biospecimens and the abcd study: rationale, methods of collection, measurement and early data. Dev. Cognit. Neurosci. 32, 97–106.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: multivariate imputation by chained equations in r. J. Stat. Softw. 45, 1–67.

VanderWeele, T.J., Ding, P., 2017. Sensitivity analysis in observational research: introducing the e-value. Ann. Intern. Med. 167, 268–274.

VanderWeele, T.J., Shpitser, I., 2013. On the definition of a confounder. Ann. Stat. 41, 196.

Volkow, N.D., et al., 2018. The conception of the abcd study: from substance use to a broad nih collaboration. Dev. Cognit. Neurosci. 32, 4–7.

Walum, H., Waldman, I.D., Young, L.J., 2016. Statistical and methodological considerations for the interpretation of intranasal oxytocin studies. Biol. Psychiatry 79, 251–257.

Wasserstein, R.L. & Lazar, N.A. The asa statement on p-values: context, process, and purpose. (2016).

Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J., Visscher, P.M., 2018. Common disease is more complex than implied by the core gene omnigenic model. Cell 173, 1573–1580.

Zucker, R.A., et al., 2018. Assessment of culture and environment in the adolescent brain and cognitive development study: rationale, description of measures, and early data. Dev. Cognit. Neurosci. 32, 107–120.

Further Reading

Appelbaum, M., et al., 2018. Journal article reporting standards for quantitative research in psychology: the apa publications and communications board task force report. Am. Psychol. 73, 3.

Boker, S., et al., 2011. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76, 306–317.

Cinelli, C., Hazlett, C., 2020. Making sense of sensitivity: extending omitted variable bias. J. R. Stat. Soc.: Series B (Statistical Methodology) 82, 39–67.

Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10, 507–521.

Jin, Y., et al., 2018. Does the medical literature remain inadequately described despite having reporting guidelines for 21 years?–a systematic review of reviews: an update. J. Multidiscip. Healthc. 11, 495.

Preacher, K.J., 2014. Extreme groups designs. Encycl. Clin. Psychol. 1–4.

Preacher, K.J., Rucker, D.D., MacCallum, R.C., Nicewander, W.A., 2005. Use of the extreme groups approach: a critical reexamination and new recommendations. Psychol. Methods 10, 178.

Slotkin, J., et al., 2012. NIH toolbox. Technical Manual.[Google Scholar].

Silberzahn, R., et al., 2018. Many analysts, one data set: making transparent how variations in analytic choices affect results. Adv. Methods Pract. Psychol. Sci. 1, 337–356.