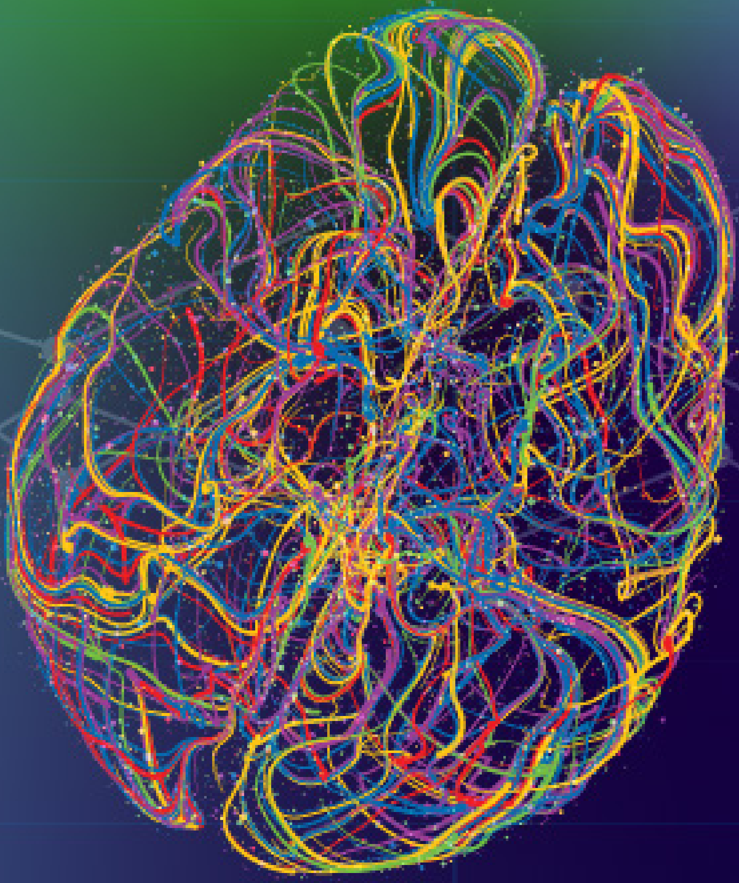
Beyond Statistical Significance: FINDING MEANINGFUL EFFECTS



VIRTUAL MEETING SUMMARY

Wednesday, September 2, 2020

NIH National Institute on Drug Abuse
Advancing Addiction Science

National Institutes of Health (NIH)
National Institute on Drug Abuse (NIDA)

Beyond Statistical Significance: Finding Meaningful Effects

Wednesday, September 2, 2020
Virtual Meeting (10:00 a.m.–6:00 p.m. EDT)

Contents

Welcome and Opening Remarks

Elizabeth A. Hoffman, Ph.D., Scientific Program Manager, Adolescent Brain Cognitive Development^{SM} Study, NIDA

Dr. Elizabeth Hoffman welcomed 653 participants to the meeting. She remarked that the high interest in the meeting was a testament to the level of engagement with the issues covered. The emergence of population neuroscience has provided unprecedented opportunities for leveraging interdisciplinary expertise to understand behavior (e.g., the interaction of environment and biology) and to develop mechanistic models to explain it. This work—including the NIH Adolescent Brain Cognitive Development (ABCD) study—involves large, heterogeneous samples of participants, which result in increased statistical significance and decreased effect size. Large cohort studies can reliably detect even small, non-null associations. Researchers must then face the question: *How do we know when a small effect is meaningful?* How researchers approach this question has statistical, clinical, biological, and public policy implications.

After reviewing the meeting agenda, Dr. Hoffman stated the objective of the meeting:

> To develop best practice recommendations for identifying, analyzing, and interpreting meaningful effects by engaging researchers from a range of disciplines in discussions of meaningful science that go beyond statistical significance.

She explained that the event would cover the big ideas underlying the objective first, then present a series of multidisciplinary concept-setting talks. This would be followed by focused presentations on three areas: (1) small effect sizes, (2) covariates/collinearity, and (3) exploratory and confirmatory frameworks. She outlined that attendees would participate in three breakout sessions, with discussions led by panelists and concept presenters with expertise in the topic and facilitated by NIH staff members. After reports from each breakout session, all participants would discuss the main issues raised during the meeting. Dr. Hoffman referred participants to the resources located on the meeting's website, including an article on meaningful effects by researchers from the ABCD Study.

Directors' Remarks

Nora D. Volkow, M.D., Director, National Institute on Drug Abuse

Dr. Nora D. Volkow welcomed participants and thanked colleagues at the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute of Mental Health (NIMH) for their contributions to the meeting. She commented on the importance of reproducibility in science. With the large data sets associated with major NIH studies, such as ABCD and the Human Connectome Project, researchers observe multiple effects that reach statistical significance. However, the meaning of these findings is not straightforward, requires consideration, and depends on the scientific question. In genetics, for example, scientists need to identify the genes involved in neurodevelopment and how they interact with the environment to influence behavior. Because the brain is highly complex, scientists are searching for the ability to analyze complexity. Psychiatric diseases are not easily identified as originating in any particular brain region. Rather, they originate from the interaction of brain processes in ways that are dynamic across development. This perspective is needed to determine, for example, why an environmental exposure affects behavior in one individual but not another, despite shared genetics. Models and paradigms that enable reproducibility are crucial, as researchers can divide up a large sample for discovery and replication samples. Models can help ensure that findings are robust and then reproduce them in other samples. Careful consideration of these issues among researchers will position science to use big data in ways that advance knowledge.

George Koob, Ph.D., Director, National Institute on Alcohol Abuse and Alcoholism

Dr. George Koob welcomed participants and commented on the tremendous changes to neuroscience in a brief period. Collaborations using big data offer an unprecedented opportunity to understand the complexity in the brain and in behavior. He stressed the importance of scientific rigor and reproducibility, particularly in the analysis and interpretation of data. NIAAA and all NIH Institutes involved in the ABCD study emphasize that the accurate interpretation of findings has implications for scientific advances and clinical applications. The large number of meeting participants shows the field's priority on and commitment to rigorous science. He suggested that each participant consider applying one idea from the meeting to his or her work.

Joshua Gordon, M.D., Ph.D., Director, National Institute of Mental Health

Dr. Joshua Gordon expressed his gratitude to participants for taking time to attend and remarked on the complexity of the brain disorders addressed by NIDA, NIAAA, and NIMH. These disorders are characterized by complex underlying biology and psychology, heterogeneity within disorders and in symptoms, and comorbidity. Researchers look for associations among disorder labels, symptoms, biology, and outcomes. The complexity of these disorders is characterized by multiple effects and some shared neural circuits. Researchers have identified various factors associated with them, and each may have small effects. The field acknowledges that genetic complexity is only one level of the overall biological intricacy. Gene expression, signaling pathways, and neural circuits (both microcircuits and macrocircuits) are engaged in multiple behaviors—resulting in complexity. With larger data sets, researchers can better understand environmental complexity. Scientists have some sense of the scope of the connectome. However, the landscape of environmental factors that contribute to these disorders is really complex. Examining associations and small but meaningful effect sizes should help shape knowledge in that important area.

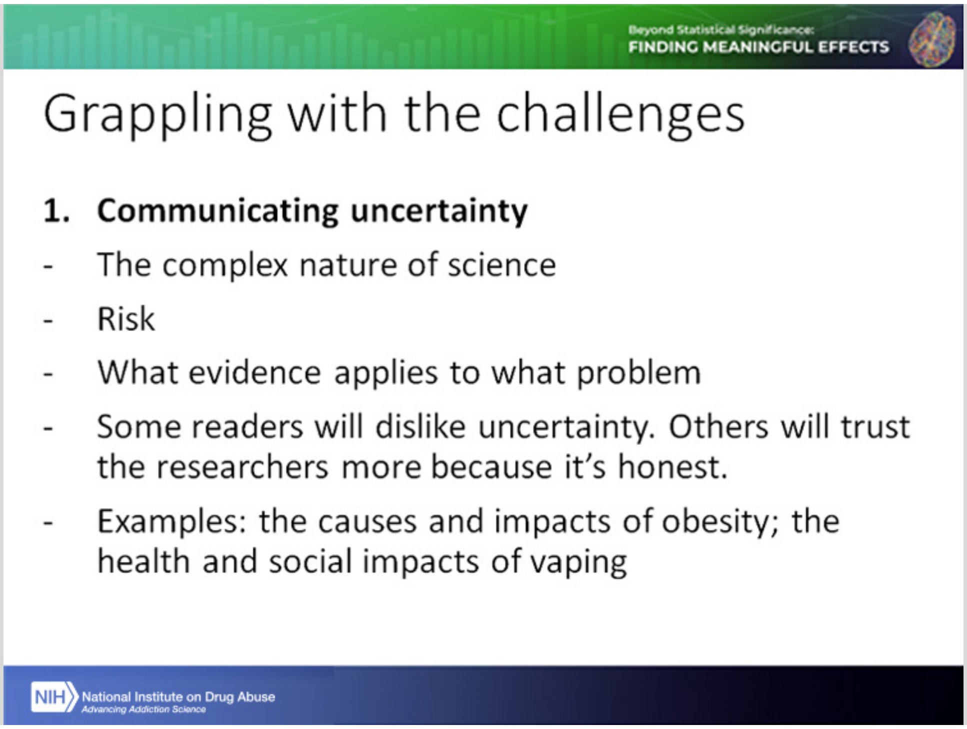
Researchers can use small, robust statistical associations in two ways to improve the lives for individuals with these disorders. First, they can use data to make clinical predictions for individuals (e.g., treatment response and illness course and severity). Clinically meaningful effects can guide patient care, and this requires combining multiple small effects into large predictors. In the case of the prodrome of schizophrenia, researchers used a variety of data points and combined small associations to develop an individual-level predictor of progression to frank psychosis with 70 percent accuracy. Second, these associations can help scientists better understand the underlying biology of these conditions (genetics, neurocircuitry) and biologically characterize groups of individuals for the purposes of developing better treatments. Findings with small effect sizes may point to overlapping biology and guide future research. NIMH supports research pursuing these approaches. In addition to the prodrome for schizophrenia, work focuses on neuroimaging predictors of treatment for obsessive compulsive disorder. The biological overlap seen between groups is not an individual predictor, but will have meaning moving forward as researchers combine the results of multiple studies.

Keynote Address—Putting the Reader First: How to Communicate Meaning in a Meaningful Way

Jessica Wapner
Science Journalist

Ms. Wapner reviewed the underlying difficulties involved in communicating science, including complexities in information, people, and context. Currently, the world is contentious, and it can be difficult to find "clean air above the fog." Yet, it is important for scientists to discuss their work in ways that are accessible to laypeople to share the excitement, expand public knowledge, correct misinformation, influence policy, and honor taxpayer-funded research. The need to correct misinformation is significant and requires patience from scientists. It is disheartening when scientists who receive taxpayer-funded grants do not engage with communicating research to the public. Lay readers of science care about cost (although this is the most common omission in news stories), benefit, possible harms, what is known, and the relevant conflicts of interest. For clinical research, lay readers focus on the seriousness of the condition and the already-available treatments.

After reviewing some specific examples of poor science reporting and why they were problematic (e.g., hype, misleading information, and anecdote rather than data), Ms. Wapner outlined a positive instance and the underlying reasons. Good science reporting is appropriately cautious and does not avoid challenges to the results. Although clickbait (defined by Oxford as "Content whose main purpose is to attract attention and encourage visitors to click on the link to a particular web page") is often derided, it is not necessarily bad. It is important for research to garner attention, but problems can arise when news stories use language that scientists feel misrepresents their work.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Grappling with the challenges

1. **Communicating uncertainty**

- The complex nature of science
- Risk
- What evidence applies to what problem
- Some readers will dislike uncertainty. Others will trust the researchers more because it's honest.
- Examples: the causes and impacts of obesity; the health and social impacts of vaping

NIH National Institute on Drug Abuse
Advancing Addiction Science

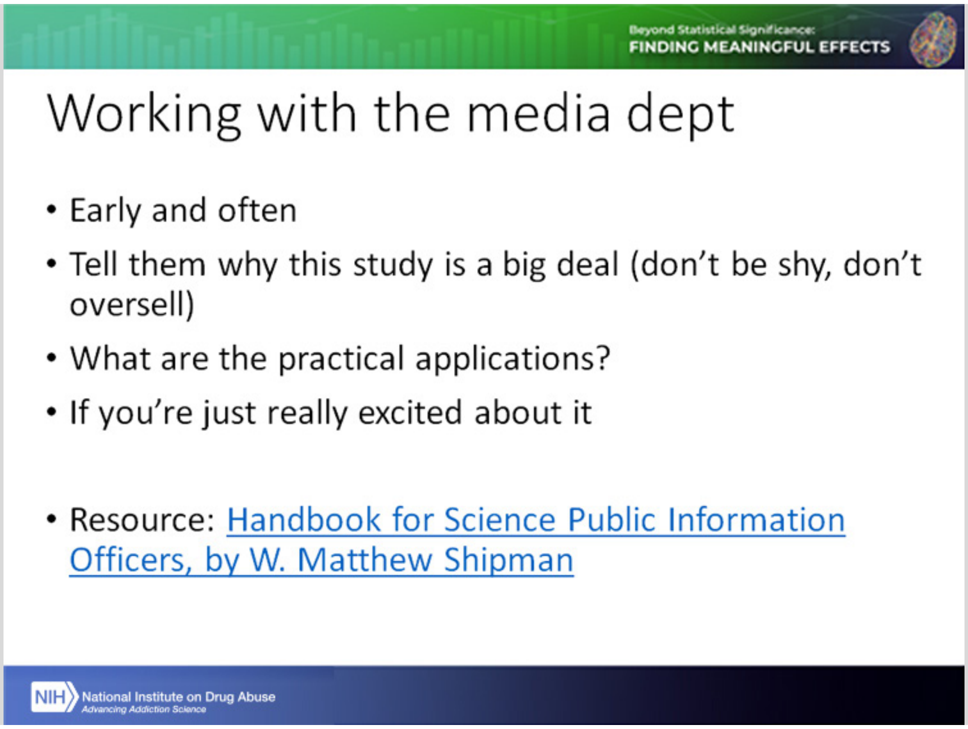
Key challenges include the difficulties of communicating uncertainty, as science is complex. Risk is a particularly difficult concept to convey. And it can be challenging to determine what evidence applies to a particular problem. Some readers will dislike uncertainty, while others will trust the researchers more because a nuanced presentation is honest. A major challenge is reader heterogeneity—including variation in prior knowledge of science, numeracy, and ways of interpreting new information. Those who want to communicate science to the general public must consider that only one in 10 Americans has had a formal education as a scientist or engineer. Additionally, evidence shows that people's background and values and beliefs shape the link between their knowledge of science and their attitude toward it. The public may not recognize that having an opinion is not equivalent to having knowledge.

To address these challenges, it is crucial for individuals who communicate science (in press releases and news stories) to always use numbers, reduce the effort needed to understand them, and explain what they mean. Science communicators must not rely on inference and should highlight important information. For example, it is best to report absolute rather than relative risk, and percentages should be accompanied by numbers or not used at all. Framing provides context to a study and is another important way to address challenges in science communication. Framing behaviors (e.g., safe sex, vaccination) in terms of gain or loss encourages or discourages behavior. Emphasis framing describes what is at stake in a complex debate and why the issue matters (e.g., a story on vaccination showing how an unvaccinated child suffered). Communicating to policymakers is a particular challenge and can be a very murky endeavor. Some scientists put in the effort because they are motivated to change policy. Such communication can occur in one direction (e.g., briefs, websites) or via "brokers" (e.g., think tanks, policy centers, nonprofits). It is often useful to present new scientific findings as "complements" to the expertise of policymakers.

Integrity, dependability, and competence engender trust in science communication. Such trust varies with political ideology, race, ethnicity, income, religiosity, education, knowledge, and levels of social capital. Sometimes people can confuse the message and the messenger. Generally, we trust those who share our interests and those whom we perceive as having expertise—which is not the same as actually having expertise. Emotions are part of science communication and help convey the meaning of the work. Strong narratives create a pathway for readers to understand and be interested in the science.

Scientists should understand the challenges faced by journalists—including tight deadlines, the need for conflict/drama, short attention spans of readers, and a lack of job security. Good reporters ask good questions about scientific methods, not only results. They also ask about the bigger picture, next steps in the research, and cautions and limitations. Good reporters may also ask about the scientist depending on the story, as well as surprising aspects of the research. They always ask an outside expert for comment on the research for balance, and this legitimizes the findings and interpretation. Scientists should make time to speak with reporters and realize that they may not be quoted. Reporters do not write headlines—editors do. Headlines are often shaped by topics of Internet searches or efforts to make a compelling story in a short word count. Scientists should not be deterred from speaking with journalists because of headlines. It is not always possible for the reporter to allow the scientist to review quotes. Scientists should remember that good reporters are not cheerleaders.

Social media can be a positive tool for scientists to disseminate messages, get out of their silos, convey research accurately, and highlight important research. Similarly, social networks are useful for scientists. Scientists should work with their institutions' media departments early and often, and explain why the work is important, particularly its practical applications and why it is exciting. Ms. Wapner acknowledged the *Handbook for Science Public Information Officers* by W. Matthew Shipman. In working with media departments, it is important for scientists to understand that press releases need to focus on a limited number (up to three) of key findings and that plain language is vital. Communicating science benefits from a compelling narrative, which does not compromise accuracy. Key information that the media department needs to know is whether the work is done in an animal model. When working with members of the media, do not imply that an observational study showed causation. Finally, if the research involves testing, explain both the specificity of the test and the sensitivity of the test (per Mr. Shipman). To make science communication meaningful, investigators should be clear about why the public should know about this work and why it is important. Scientists should clarify the angles/takes that are inappropriate, convey the significance of the work, and identify the pieces of information essential to conveying the meaning of the study.


Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS
NIH National Institute on Drug Abuse
Advancing Addiction Science

Working with the media dept

- Early and often
- Tell them why this study is a big deal (don't be shy, don't oversell)
- What are the practical applications?
- If you're just really excited about it
- Resource: Handbook for Science Public Information Officers, by W. Matthew Shipman

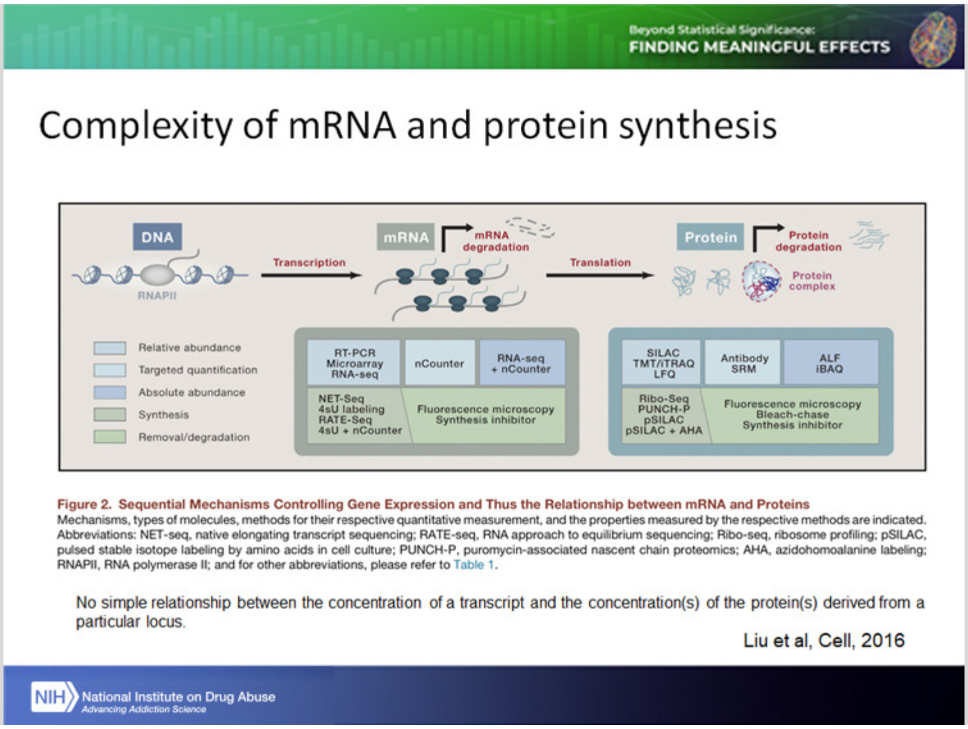
Dr. Hoffman thanked Ms. Wapner for her presentation and focused on the collective responsibility to communicate science effectively.

Concept Presentations

Meaningful Effects in Single Cell Transcriptomics and Epigenetics Data

Mike Hawrylycz, Ph.D.
Allen Institute for Brain Science

Dr. Hawrylycz focused on reproducibility of research in genetics and explained that non-reproducibility of results is a major problem in biomedical research in general—with rates as high as 65 percent to 89 percent in pharmacological studies and 64 percent in psychological studies. The problem is especially prominent in high-dimensional experiments, such as gene expression analyses in which thousands of hypotheses are being tested simultaneously. Even with strict multiple-hypothesis correction, it has been shown that about 72 percent of genomic associations initially reported as significant were found to be an overestimate of the true effect when subsequently tested in additional data sets. Both biological models and analytic methods contribute to irreproducibility in biomedical research. Analyses focus on significance (p-values) rather than effect size or independent verification. A problem with reporting significance based on p-values is that researchers do not necessarily understand the underlying data distribution and assumptions of normality confound the work. Instead, rigorous experimental design should focus on incorporating and accounting for study heterogeneity as appropriate and then validating results in independent cohorts.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Complexity of mRNA and protein synthesis

Figure 2. Sequential Mechanisms Controlling Gene Expression and Thus the Relationship between mRNA and Proteins
Mechanisms, types of molecules, methods for their respective quantitative measurement, and the properties measured by the respective methods are indicated. Abbreviations: NET-seq, native elongating transcript sequencing; RATE-seq, RNA approach to equilibrium sequencing; Ribo-seq, ribosome profiling; pSILAC, pulsed stable isotope labeling by amino acids in cell culture; PUNCH-P, puromycin-associated nascent chain proteomics; AHA, azidohomoalanine labeling; RNAPII, RNA polymerase II; and for other abbreviations, please refer to Table 1.

No simple relationship between the concentration of a transcript and the concentration(s) of the protein(s) derived from a particular locus.

Liu et al, Cell, 2016

NIH National Institute on Drug Abuse
Advancing Addiction Science

The synthesis of mRNA and proteins is a complex process, and much is not yet understood about the end results. To measure gene expression quantitatively and analyze the results, researchers have several different methods to choose from (e.g., Northern blotting, *in situ* hybridization, microarrays, serial analysis of gene expression, RNA-seq, and spatial transcriptomics). Thousands of microarray studies published annually are potentially confounded by “batch effects,” the systematic error introduced when samples are processed in multiple batches. Researchers have a tendency to ignore past results and revisit the question with new technologies. Although batch effects can be reduced by careful experimental design, they cannot be eliminated unless the whole study is done in a single batch. Tools to address this include ComBat and empirical Bayes.

The hybridization-based RNA-seq experimental technique performs poorly on absolute transcript quantification. RNA-seq allows for actual counting of RNA molecules—for example, by counting the sequenced fragments per kilobase of exon model per million mapped reads (FPKM). Only a fraction of all transcripts in a sample are being sequenced, and the observed read counts, such as FPKM, are biased by the sequence of the transcript. Thus, absolute quantification requires additional calibration of the data.

Spatial molecular mapping is a very important technique to describe cell classes and can help researchers understand brain transcription. Epigenetic changes can be measured with chromatin accessibility and conformation assays, ATAC-seq, and DNase I hypersensitivity. Researchers can also conduct a DNA methylation analysis or an analysis of DNA/protein interactions, which is often studied via chromatin immunoprecipitation. The most common purpose of a gene expression study is to find statistically differentially expressed genes (DEGs), which are determined by comparing sample-level gene expression data between cases and controls (differential expression analysis). Researchers can use DEGs to gain insight into disease pathophysiology, to serve as clinical biomarkers, and to target for pharmacologic therapy. It is common to claim a condition-specific association for a gene (e.g., define a novel tissue–specific marker or associate a gene with a particular perturbation). Technical reproducibility is an essential but far from sufficient prerequisite for clinical use of any novel diagnostic test. A more important issue is the accuracy of result prediction and its clinical value.

Single-cell RNA sequencing (scRNA-seq) has developed rapidly as a powerful technology for studying transcriptomics at the single-cell level. However, compared with bulk-cell data, scRNA-seq data has a higher level of noise due to both biological and technical reasons (e.g., lower-input materials, cell-cycle phase, amplification biases, and the dropout and bursting events). Such events are caused by the stochastic nature of the gene expression process at the single-cell level. The dropout events generate zero expression, statistically leading to zero inflation in the gene-expression distribution at a much higher proportion than expected under the standard negative binomial model commonly assumed in bulk-cell data.

Functional misinterpretation of RNA-sequencing (RNA-seq) data can occur. Data normalization—aiming to remove systematic effects from the data to ensure that technical biases have minimal impact on the results—is a critical step in RNA sequencing (RNA-seq) analysis. Analyzing numerous RNA-seq data sets, researchers detected a prevalent sample-specific length effect that leads to a strong association between gene length and fold-change estimates between samples. This stochastic sample-specific effect is not corrected by common normalization methods, including reads per kilobase of transcript length per million reads, Trimmed Mean of M values, relative log expression, and quantile and upper-quartile normalization. Importantly, this bias causes recurrent false positive calls by gene-set enrichment analysis methods, thereby leading to frequent functional misinterpretation of the data. Gene sets characterized by markedly short genes (e.g., ribosomal protein genes) or long genes (e.g., extracellular matrix genes) are particularly prone to such false calls.

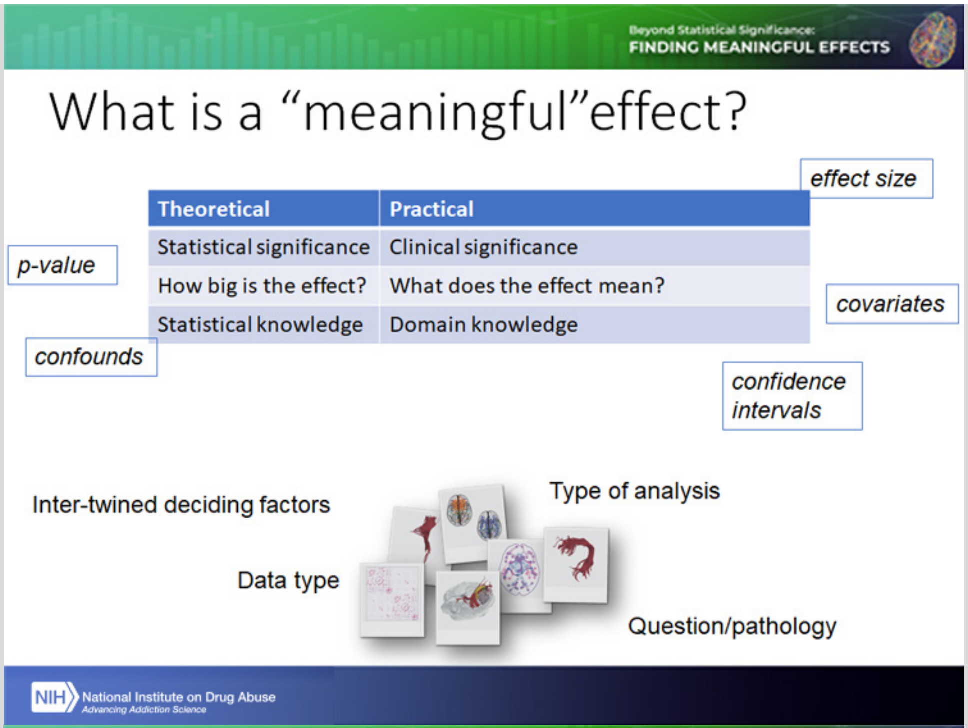
Through single-cell characterization (transcriptomic, physiological, and morphological) by the use of genetic tools followed by data-driven classification and mapping connectivity from defined cell types, researchers can determine the role of various cell types in neural circuit function. When characterizing the cell-type parameter space, researchers must ask: *How do experimental factors affect the final cell type classification? How do computational factors affect the final cell type classification? How do biological factors affect the final cell type classification?* The genetic toolkit for the analysis of cell types includes special mouse models and techniques for identifying genes and enhancers, validating transgenic lines, and characterizing cell types. Tools also exist for visualizing cell-type clusters (e.g., UMAP and t-SNE). Although these tools offer sophisticated ways to present information, they are not quantitatively faithful and may not illuminate neural processes. They provide a signal, but researchers do not yet understand the natural bounds and variability of this data. Heterogeneity may be the rule.

Approaches to identifying meaningful effects for cell types include: (1) statistical measures to define cell type and topological methods from discrete Morse theory; (2) coherence across profiling techniques and platforms, and agreement in cell type assignments across distinct RNA-seq platforms; (3) correspondence of multimodal data sets, such that concordant transcriptomic and epigenomic results reveal cell type-specific gene regulatory landscape; (4) conservation of evolutionary and comparative genomics, which allow alignment of homologous cell types and definition of a consensus taxonomy; (5) cell-type nomenclature and ontology, including Neuron Phenotype Ontology, which offers model neurons as extensible collections of phenotypic properties; and (6) architecture for community sharing of data, such as knowledge graphs for biomedical data (as in the NIH Brain Research through Advancing Innovative Neurotechnologies® Initiative).

Modeling Meaningful Effects in Neuroimaging Studies

Ragini Verma, Ph.D.
University of Pennsylvania

Dr. Verma focused on modeling meaningful effects in neuroimaging studies and began with a characterization of meaningful effects that includes theoretical (statistical significance [p-value] and knowledge [confounds]) as well



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

What is a "meaningful"effect?

Theoretical	Practical
Statistical significance	Clinical significance
How big is the effect?	What does the effect mean?
Statistical knowledge	Domain knowledge

effect size

p-value

covariates

confounds

confidence intervals

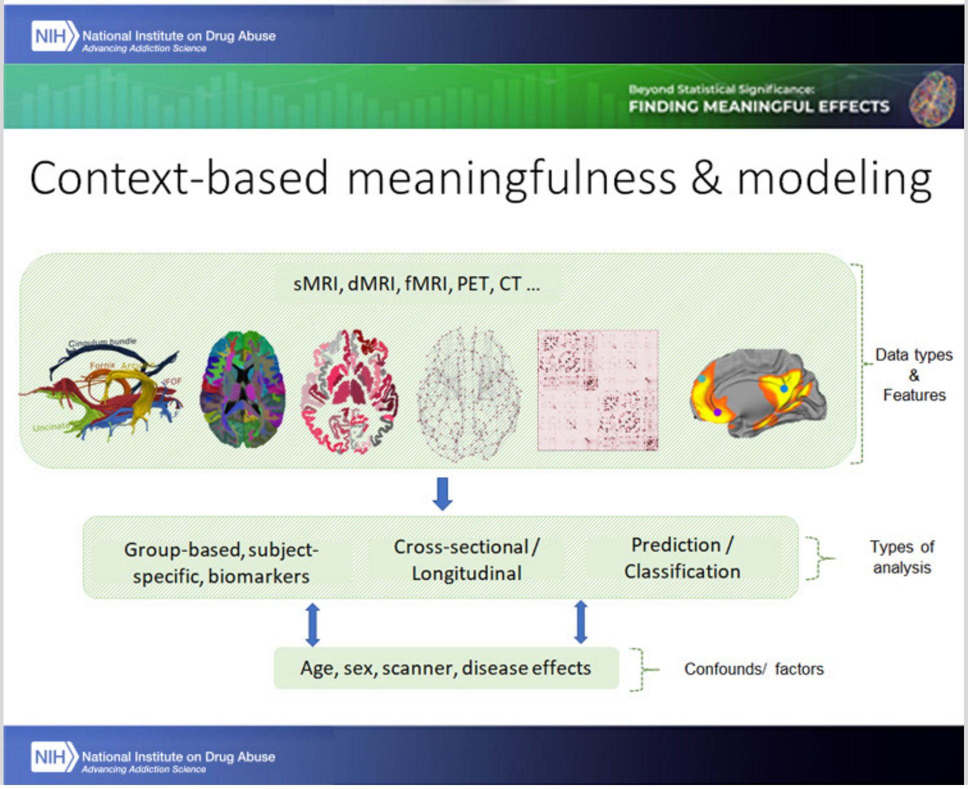
Inter-twined deciding factors

Type of analysis

Data type

Question/pathology

NIH National Institute on Drug Abuse
Advancing Addiction Science



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Context-based meaningfulness & modeling

sMRI, dMRI, fMRI, PET, CT ...

Data types & Features

Group-based, subject-specific, biomarkers

Cross-sectional / Longitudinal

Prediction / Classification

Types of analysis

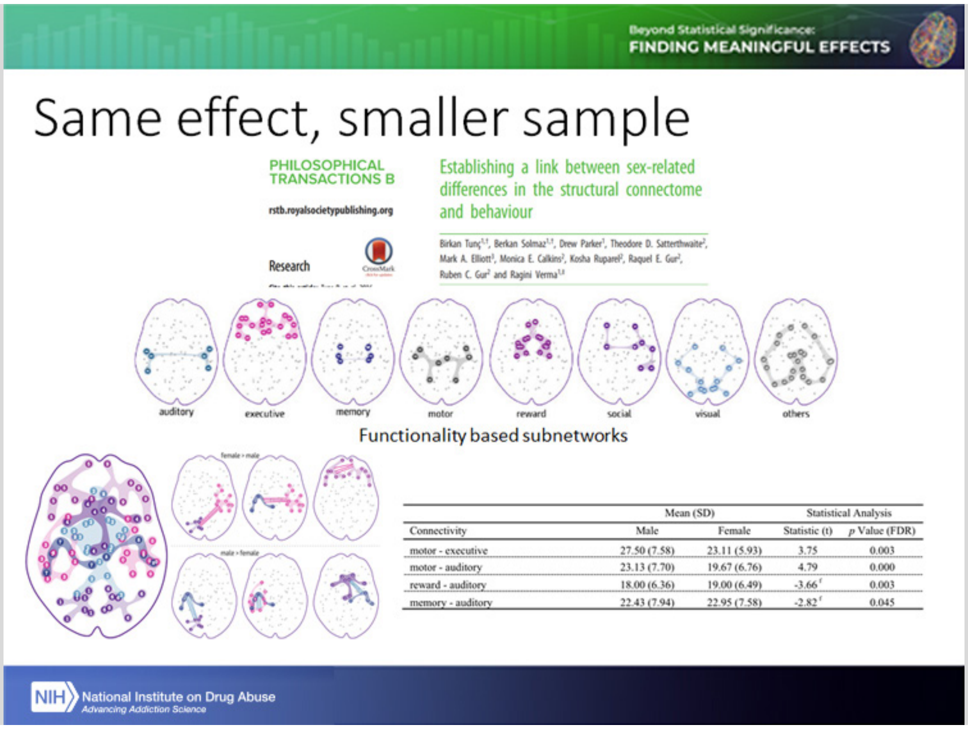
Age, sex, scanner, disease effects

Confounds/ factors

NIH National Institute on Drug Abuse
Advancing Addiction Science

as practical (clinical significance [effect size] and domain knowledge [covariates]) considerations. Other relevant considerations include intertwined deciding factors, the type of data and analysis, and the underlying question or pathology of interest.

Context-based meaningfulness and modeling processing differences involves particular data types (sMRI, dMRI, fMRI, PET, or CT, etc.) and features, types of analyses (e.g., group-based or subject-specific, cross-sectional or longitudinal, and prediction or classification), and confounds/factors (e.g., age, sex, scanner, and disease effects). With biological and pathological group differences, researchers must account for variation and longitudinal changes, and consider clinical significance and big data integration.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Same effect, smaller sample

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Establishing a link between sex-related differences in the structural connectome and behaviour

Birkan Tunç[1,†], Berkan Solmaz[1,†], Drew Parker[1], Theodore D. Satterthwaite[2], Mark A. Elliott[3], Monica E. Calkins[2], Kosha Ruparel[2], Raquel E. Gur[2], Ruben C. Gur[2] and Ragini Verma[1,‡]

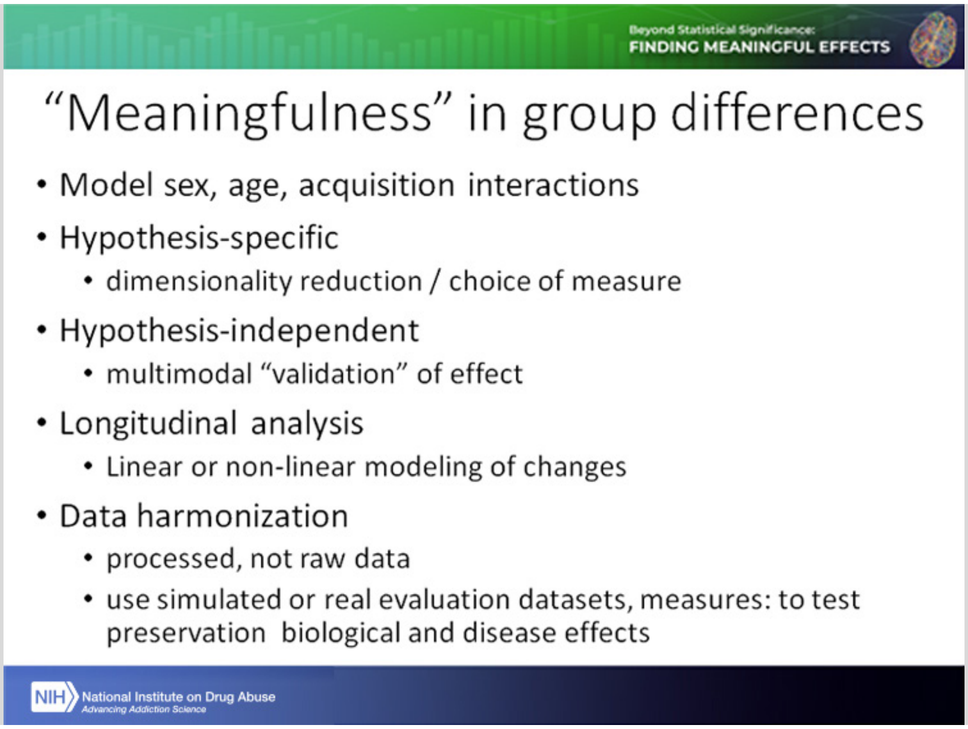
Functionality based subnetworks

	Mean (SD)		Statistical Analysis	
Connectivity	Male	Female	Statistic (t)	*p* Value (FDR)
motor - executive	27.50 (7.58)	23.11 (5.93)	3.75	0.003
motor - auditory	23.13 (7.70)	19.67 (6.76)	4.79	0.000
reward - auditory	18.00 (6.36)	19.00 (6.49)	-3.66[f]	0.003
memory - auditory	22.43 (7.94)	22.95 (7.58)	-2.82[f]	0.045

NIH National Institute on Drug Abuse
Advancing Addiction Science

Dr. Verma provided the example of sex differences in the structural connectome found among 949 subjects with no psychopathology from the Philadelphia Neurodevelopmental Cohort. The same effects were found in a smaller sample in a study that examined a disease-related (traumatic brain injury [TBI]) group difference, and the question arose: *Given connectivity differences in TBI, should researchers control for sex?* Controlling for age-sex as a covariate may not be enough clinically, and effect size analysis is crucial. There was a need to investigate the sexes separately. In moderate-severe TBI patients, researchers observed longitudinal changes. Cross-sectional analysis at each time point showed a significant difference. Researchers needed to determine whether a linear model would show the observed longitudinal changes, and they controlled for sex, age, and time since injury. They found that patient outcome was predicted equally by the imaging score and the clinical measure of injury. The clinical injury score was more expensive to acquire and not consistent across sites. Therefore, imaging results were better for outcome prediction.

A second example examined whether researchers should hypothesize or not regarding group differences in big data. Significant differences are often observed with small sample size, and in this case, the confidence interval and effect size are crucial. Group differences in big data are an issue in a wide range of areas, such as autism, TBI, psychopathology, development, and life span studies. Such effects allow researchers to parse heterogeneity of the disease spectrum and improve power to characterize the effects of disease, age, and sex. They also offer researchers opportunities to combine controls across studies, explore new hypotheses, and identify areas where acquisition will contribute most in future. When researchers observe that the biggest effect is associated with acquisition data and processing differences, big data can be very messy. The simple solution is to use the scanner as a covariate. However, a number of issues come into play. These include that scanner differences may not be linear across the brain or across measures. Also, different measures may vary in how they manifest the effect. Although researchers can remove scanner differences, they cannot eliminate the effects of pathology and biology.



Beyond Statistical Significance: FINDING MEANINGFUL EFFECTS

“Meaningfulness” in group differences

- Model sex, age, acquisition interactions
- Hypothesis-specific
  - dimensionality reduction / choice of measure
- Hypothesis-independent
  - multimodal “validation” of effect
- Longitudinal analysis
  - Linear or non-linear modeling of changes
- Data harmonization
  - processed, not raw data
  - use simulated or real evaluation datasets, measures: to test preservation biological and disease effects

NIH National Institute on Drug Abuse
Advancing Addiction Science

To determine “meaningfulness” in group differences, researchers should model sex, age, and acquisition interactions. They can take a hypothesis-specific approach through dimensionality reduction and choice of measure. They can also take a hypothesis-independent approach through multimodal “validation” of an effect. Investigators can also use a longitudinal analysis with linear or nonlinear modeling of changes. Finally, they can determine “meaningfulness” by harmonizing data—using processed (not raw) data and simulated or real evaluation data sets, with measures to test the preservation of biological and disease effects.

Subject-wise measures are aligned with precision medicine. Here, the questions about effects are: *How is a patient different from the population? How does a new patient relate to known cases?* Ideally, researchers would be able to parse the heterogeneity, make prognostic and diagnostic predictions, and decide on outcomes and treatments. Realistically, researchers are modeling the “healthy” variation, work with too many univariate measures, and are subject to unconscious bias in training their models. Dr. Verma reviewed examples of effects indicating how different the patient is from a healthy individual, treatment differences, and subject-specific disease markers.

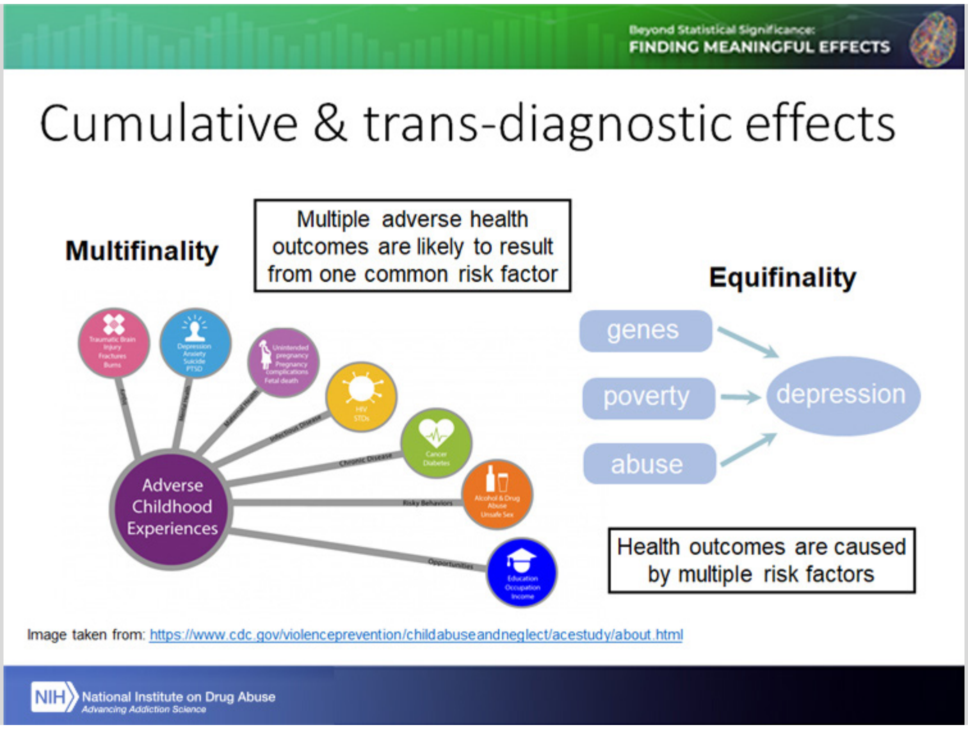
“Meaningfulness” of effects in precision medicine can be outlier detection, in which researchers model the space of healthy controls to capture the variation attributable to sex, age, education, and other variables and identify the factors that are important to the clinical cohort. For multivariate markers, less is more. “Meaningfulness” of effects can also be biomarker creation. For this activity, it is important to balance the training sample (e.g., sex, age, and education). Accuracy of the model is not enough, as a sensitivity/specificity analysis is key. Researchers must also correct for confounds *post hoc*. Dr. Verma stressed that clinical significance outweighs statistical significance, domain knowledge is crucial in assessing relevance, and cross-validation across data sets is key.

Exploratory, Confirmatory Frameworks—Identifying Meaningful Effects at the Intersection Between Genes, Life Experiences, and Development

Erin Dunn, Sc.D., M.P.H.
Massachusetts General Hospital

When one considers the definition of a “meaningful” effect, p-values, confidence intervals, and statistical and clinical significance all come to mind. Other issues include the question of “who decides what is a ‘meaningful’ effect?” In answering this question, researchers must balance incremental and high-risk science. They must also obtain results in ways that enable scientists to build (rebuild) trust with the public. It is important to advance work on “meaningful” effects to achieve a larger social mission—that is, subgroup analyses based on socioeconomic status, race/ethnicity, and sex. Another question that emerges is, “how should we decide what is a ‘meaningful’ effect?” To answer this question, researchers must decide whether the goal is to account for 100 percent of the variability in the outcome and determine a process for creating standards without perpetuating binary thinking. The question of what is a “meaningful” effect depends on the research (short-, medium, and longer-term), intervention, and policy agendas.

In the public health approach, which involves achieving a shift to the left of the population distribution through measures that target the entire population, small effects can have a large effect. But when researchers work in silos, they can forget that variables are interrelated. There can be cumulative and trans-diagnostic effects, so that small effects add up to have a significant impact. In multifinality, multiple adverse health outcomes are likely to result from one common risk factor (e.g., adverse childhood experiences). In equifinality, health outcomes are caused by multiple risk factors (e.g., genes, poverty, and abuse all contribute to depression).



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Cumulative & trans-diagnostic effects

Multiple adverse health outcomes are likely to result from one common risk factor

Health outcomes are caused by multiple risk factors

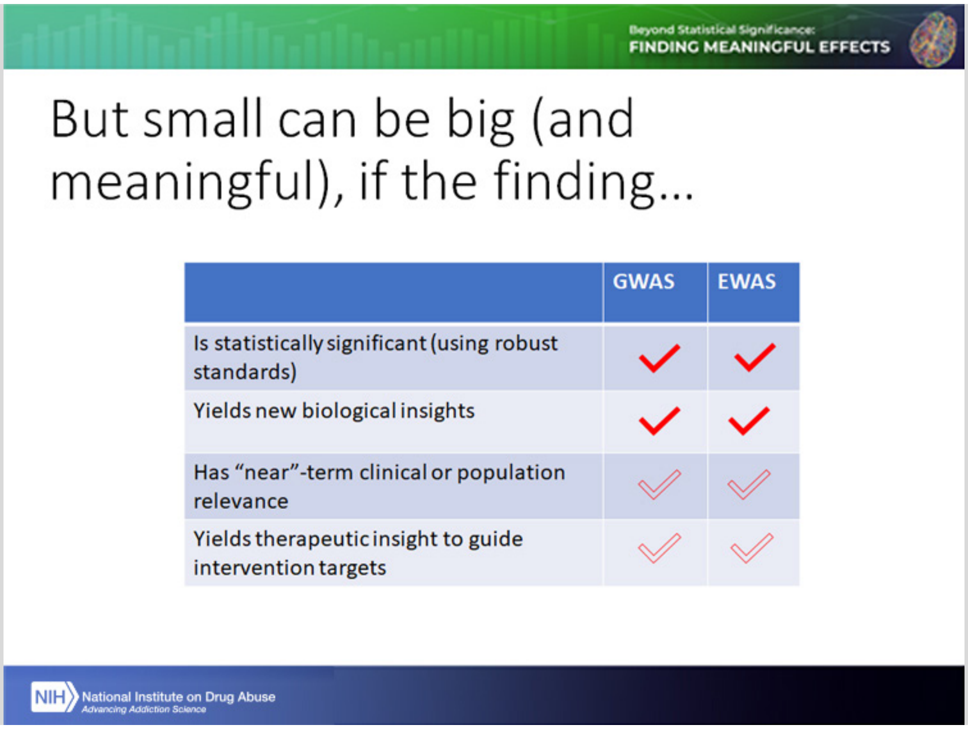
Multifinality
Traumatic Brain Injury Fractures Burns
Depression Anxiety Suicide PTSD
Unintended pregnancy Pregnancy complications Fetal death
HIV STDs
Cancer Diabetes
Alcohol & Drug Abuse Unsafe Sex
Education Occupation Income
Injury
Mental Health
Maternal Health
Infectious Disease
Chronic Disease
Risky Behaviors
Opportunities
Adverse Childhood Experiences
Equifinality
genes
poverty
abuse
depression

Image taken from: https://www.cdc.gov/violenceprevention/childabuseandneglect/acestudy/about.html

NIH National Institute on Drug Abuse
Advancing Addiction Science

Finding meaningful effects of gene-environment interplay across the life course requires consideration of five main points. First, the field has seen a massive pendulum swing from hypothesis-driven to hypothesis-free research at the intersection between genes and environments. It is important to understand that psychiatric genetics has a history of extremes. The candidate gene era had both strengths and weaknesses. Its strengths were a small number of tests of statistical association and that smaller sample sizes were acceptable. The limitations were that the science was hypothesis driven (based on presumed biological underpinnings) and that there was poor replication of candidate genes alone and in interaction with stressful life experiences. The era of genome-wide association study (GWAS) combined with a data-driven candidate gene approach was strong because it combined hypothesis-free science (i.e., GWAS) with hypothesis-driven work based on replicable loci (i.e., nonhistorical candidates). However, a limitation of this approach was the need to differentiate “good” from “bad” candidates.

The era of genome- and epigenome-wide studies (GWAS and EWAS) is characterized by hypothesis-free research in which GWAS yields more replicable results. However, a limitation is that there are a large number of statistical tests of association, so very large samples are needed. Because the search space is large, researchers are looking for “needles” (i.e., variants) in a “haystack” (i.e., the genome/epigenome). The first decade of psychiatric GWAS and EWAS has produced mostly small effects. In the largest GWAS of depression, the researchers performed a genome-wide meta-analysis that involved 807,553 subjects from three analyses (246,363 cases and 561,190 controls). The investigators sought to replicate the findings in an independent cohort (414,055 cases of self-reported clinical diagnosis of depression and 892,299 controls). They found 107 statistically independent loci that had the same direction of effect in each study and the replication sample. The small effects had a minimum odds ratio of 0.96 and a maximum odds ratio of 1.05. The study estimated that depression was 37 percent genetic and 63 percent nongenetic. Researchers can now explain about 9 percent of the variation in depression, which shows significant progress. The first and largest EWAS of depression involved a discovery cohort (7,948 European participants from nine population-based cohorts) and a replication cohort (3,308, mostly African Americans, from two population-based cohorts). The results indicated that the three CpG loci were associated with depression, with two of the discovered loci replicated in the independent sample. All three findings implicated axon guidance as the common disrupted pathway. For small effects, a one-unit difference in the Center for Epidemiologic Studies Depression Scale was associated, on average, with increases in DNA methylation by 0.03-0.05 percent for the top loci.



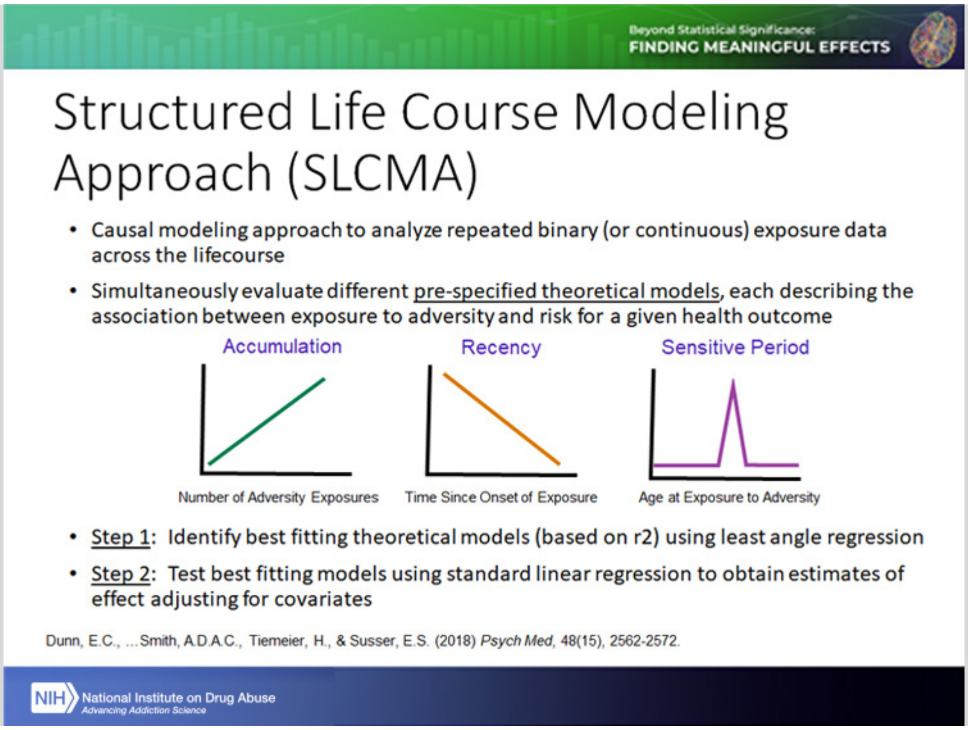
Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

But small can be big (and meaningful), if the finding...

	GWAS	EWAS
Is statistically significant (using robust standards)	✓	✓
Yields new biological insights	✓	✓
Has “near”-term clinical or population relevance	✓	✓
Yields therapeutic insight to guide intervention targets	✓	✓

NIH National Institute on Drug Abuse
Advancing Addiction Science

Dr. Dunn argued that small effects can be big (and meaningful) if the finding meets four criteria: (1) is statistically significant using robust standards, (2) yields new biological insights, (3) has “near”-term clinical or population relevance, and (4) generates therapeutic insight to guide intervention targets. Researchers continue to recognize that genetic effects are non-deterministic—genes and environment interact. The intersection between genes, life experiences, and development is a matter of “nature through nurture” rather than “nature versus nurture.” Individuals are embedded in a complex context of environment—which involves social and economic policies, institutions, neighborhoods and communities, living conditions, social relationships, and individual situations—and biology (genetic factors and pathophysiologic pathways) across the life course that affect individual and population health. Understanding this complexity requires a sophisticated data structure and analyses. Researchers must adopt new methods and adapt effective techniques to big data.


Accumulation
Recency
Sensitive Period
Number of Adversity Exposures
Time Since Onset of Exposure
Age at Exposure to Adversity

Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Structured Life Course Modeling Approach (SLCMA)

- Causal modeling approach to analyze repeated binary (or continuous) exposure data across the lifecourse
- Simultaneously evaluate different pre-specified theoretical models, each describing the association between exposure to adversity and risk for a given health outcome
- Step 1: Identify best fitting theoretical models (based on r2) using least angle regression
- Step 2: Test best fitting models using standard linear regression to obtain estimates of effect adjusting for covariates

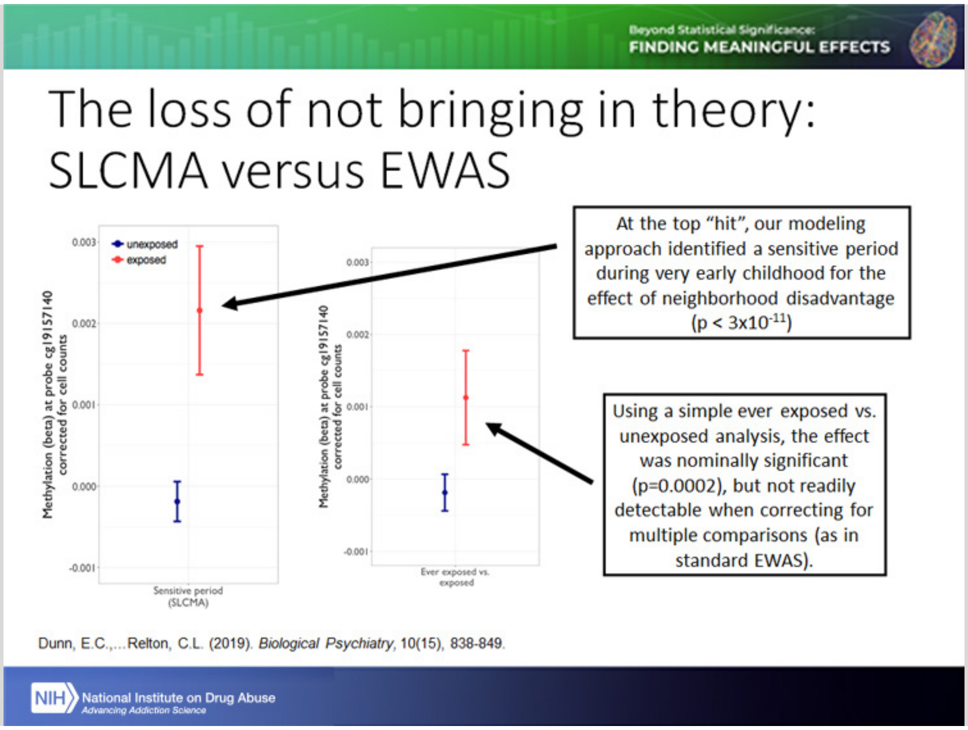
Dunn, E.C., ...Smith, A.D.A.C., Tiemeier, H., & Susser, E.S. (2018) *Psych Med*, 48(15), 2562-2572.

NIH National Institute on Drug Abuse
Advancing Addiction Science

When it comes to exploratory and confirmatory frameworks, Dr. Dunn remarked that it does not have to be one versus the other. There are new statistical approaches that allow researchers to incorporate both ways of thinking. One is the Structured Life Course Modeling Approach (SLCMA), in which researchers analyze repeated binary (or continuous) exposure data across the life course to develop a causal model. SLCMA allows investigators to evaluate simultaneously different pre-specified theoretical models, each describing the association between exposure to adversity and risk for a given health outcome. The first step is to identify the best fitting theoretical models (based on R^2) using least angle regression. The second step is to test the best fitting models using standard linear regression to obtain estimates of effect adjusting for covariates.

SLCMA was used in the Avon Longitudinal Study of Parents and Children with about 6,000 subjects. The researchers focused on childhood adversity and wanted to determine whether accumulation, recency, or sensitive period best explains the variation in outcomes. They found that the timing of adversity matters most (i.e., during the first 3 years of life). The results pointed to 38 significant loci distributed throughout the genome. Of these 38 CpG sites, more than half were identified as being influenced by exposure to adversity in very early childhood. When researchers do not incorporate theory, there can be a loss of understanding. Using SLCMA, Dr. Dunn's team identified a sensitive period during very early childhood for the effect of neighborhood disadvantage, which was highly statistically significant. In the EWAS approach, using a simple ever exposed versus unexposed analysis, this effect was only nominally significant and not readily detectable when correcting for multiple comparisons. That is, the effect would have been missed in the EWAS approach.

In the context of big data in particular, deciding which covariates to include is important. Dr. Dunn shared that a reviewer asked her team: "Why did you include baseline parent social class as a covariate? There is conceptual overlap." Addressing this question adequately required seven pages of supplemental materials (15 percent of the document), because the covariate of parent social class is complex. It was modestly correlated ($r<0.45$) with other aspects of socioeconomic status. The goal was to control for the potential confounding effects of the social class into which children are born. Team members defined confounding based on the causal inference literature. They considered the theory and empirical findings on socioeconomic status and its effects on childhood adversity and DNA methylation. The researchers also compared the results with and without adjustment for parent social class. When this factor was not adjusted for, the results were largely but not fully collapsible. The findings indicated 38 loci (same as the original findings), 31 of which were shared by the primary analysis. The betas, standard errors, and R^2 were effectively unchanged. The new analysis identified seven new sites, and seven others were no longer



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

The loss of not bringing in theory: SLCMA versus EWAS

At the top "hit", our modeling approach identified a sensitive period during very early childhood for the effect of neighborhood disadvantage ($p < 3x10^{-11}$)

Using a simple ever exposed vs. unexposed analysis, the effect was nominally significant (p=0.0002), but not readily detectable when correcting for multiple comparisons (as in standard EWAS).

Dunn, E.C.,...Relton, C.L. (2019). *Biological Psychiatry*, 10(15), 838-849.

National Institute on Drug Abuse
Advancing Addiction Science

significant when no longer adjusting. In this case, the researchers found both positive (observed association is biased away from null) and negative confounding (association is biased toward the null) of parent social class.

Discussion

Dr. Hawrylycz, can you provide more information about the use of empirical Bayes in the type of research you outlined?

Dr. Hawrylycz responded that these methods are not used as much as they should be, although some models for cell types involve their use.

Dr. Verma, is there an objective method (e.g., effect size threshold) to determine whether it is more appropriate to adjust for covariates or run separate analyses?

Dr. Verma commented that it depends on the study and particular scientific question. The issue is not cut and dried. An attendee commented that if researchers use an endpoint that says nothing about patient benefit, and find it to be statistically significant, this is not helpful for prognosis. He asked: *Isn't this a failure in a choice of measurement versus a criticism of the inference drawn from a statistical inference?* Dr. Verma agreed and emphasized that the choice of measure is very important and should be done prior to planning the analysis. In her view, measures and statistics go hand in hand, but the clinical question and impact are key.

Dr. Dunn, how can researchers empirically define weights for specific environmental exposures and what is the appropriate way to model differences in the impact of exposures?

Dr. Dunn appreciated this point and noted that researchers sometimes mistakenly view exposures as equivalent and simply sum them. For these kinds of analyses, it is important to not assume the strength of each adversity, but rather model thoughtfully with respect to various dimensions. She added that researchers are learning how to group exposures based on presumed mechanisms (e.g., deprivation versus threat as a way to group adverse childhood experiences). It is also crucial to consider the timing of exposures. An attendee asked whether Dr. Dunn's team had considered using a propensity score methodology. Dr. Dunn commented that the team is now working with this methodology.

Dr. Dunn, please elaborate on negative confounding.

If researchers remove a confounder from an analysis and see new results that are significant, it seems like evidence that the results are due to confounding. *How would one know that the confounder was inappropriately included?* Dr. Dunn remarked that the challenge is that researchers still do not know what is or is not appropriate with respect to covariates. Her team considers covariates a great deal. In hindsight, covariates now known to be heritable should not have been included in analyses (e.g., in GWAS, do not adjust for heritable covariates). Researchers learn as they go, should not pretend to know the truth, and must be transparent about covariates so others can learn.

Dr. Verma, can in vitro phantoms help us quantify scanner differences across sequence and scanner types?

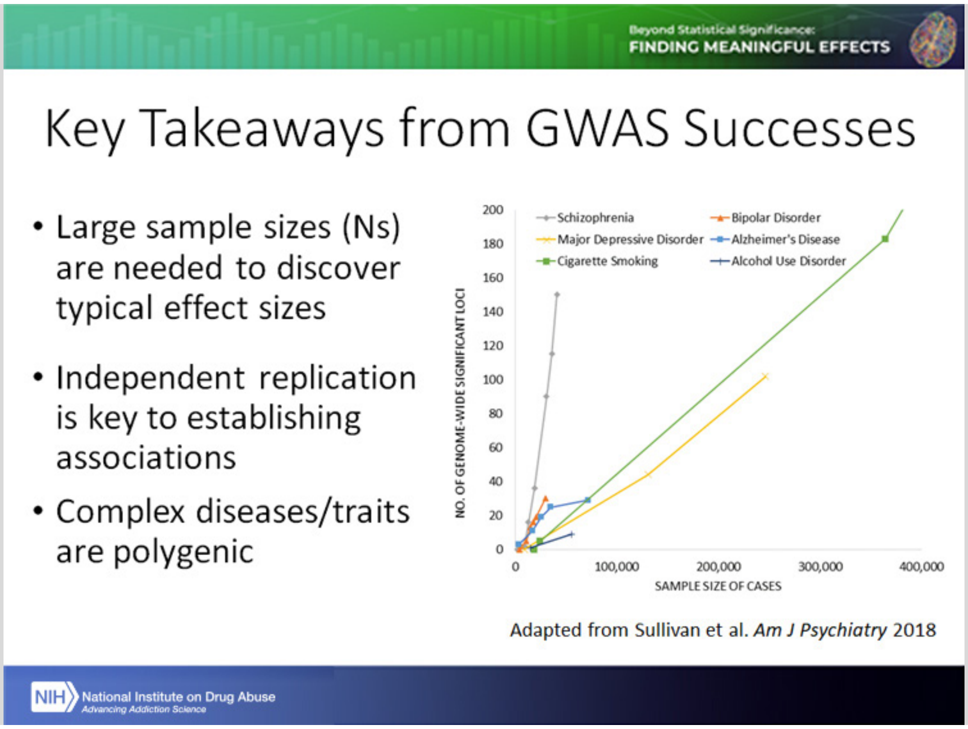
Dr. Verma replied that no phantoms currently match diffusion measurements (only gross measurements). Work is being done on developing phantoms in Pittsburgh, but this is not yet commercially available.

Panel Presentations

Small Effect Sizes—Accumulating Evidence from Small Effect Sizes: Examples in Moving From Genome-wide Association Studies to Biology and Clinical Prediction

Dana Hancock, Ph.D.
RTI International

Dr. Hancock addressed three categories of meaningful effects: statistical, biological, and clinical. For GWAS investigations that test millions of single nucleotide polymorphisms (SNPs) across the genome for association with a specified disease or trait, $p<5x10^{-8}$ quickly became field standard for statistical significance. GWAS research produced an explosion of associations. She reported that there have been more than 147,000 significant associations from more than 4,200 publications as of April 2019 (https://www.ebi.ac.uk/gwas/). As for the typical effect sizes in the GWAS Catalog, an odds ratio of 1.95 is at the 90th percentile and one of 1.22 is at the 50th percentile. Key takeaways from GWAS successes are that (1) large sample sizes are needed to discover typical effect



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Key Takeaways from GWAS Successes

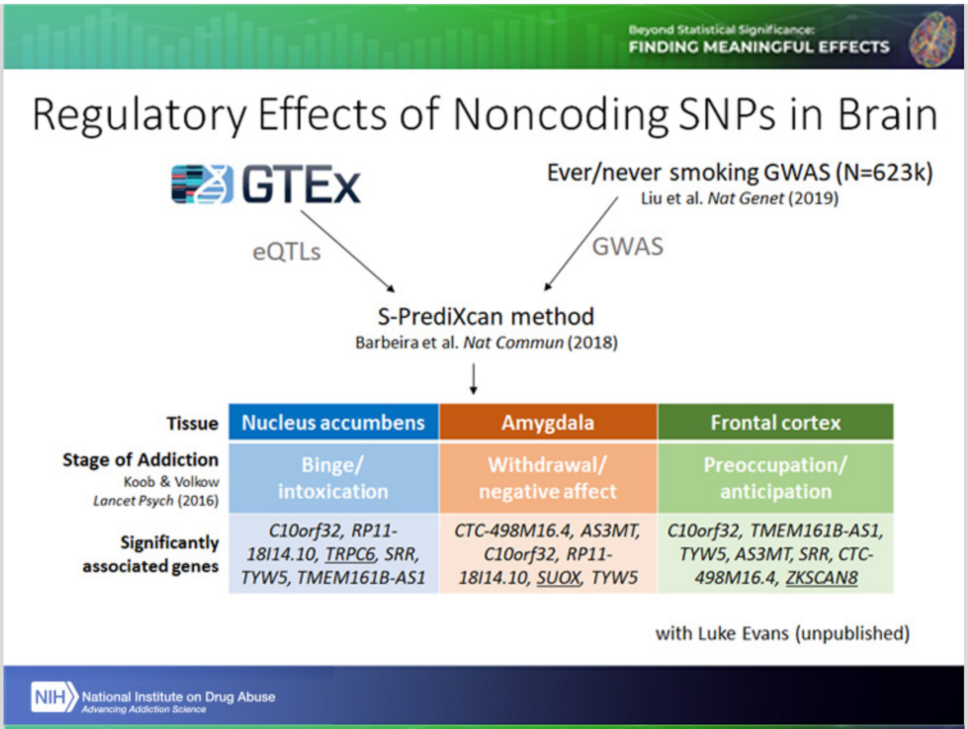
- Large sample sizes (Ns) are needed to discover typical effect sizes
- Independent replication is key to establishing associations
- Complex diseases/traits are polygenic

Adapted from Sullivan et al. *Am J Psychiatry* 2018

NIH National Institute on Drug Abuse
Advancing Addiction Science

sizes, (2) independent replication is key to establishing associations, and (3) complex diseases or traits are polygenic. Over time, replication is crucial for establishing statistically meaningful effects.

The category of biologically meaningful effects includes coding SNPs, which represent about 3.6 percent of SNPs found in the GWAS Catalog. For example, the coding SNP rs1229984 affects the protein ADH1B (alcohol dehydrogenase), which influences alcohol use disorder. Noncoding SNPs can also have biologically meaningful effects because of regulatory impacts and represent 57.2 percent of SNPs found in the GWAS Catalog. The GWAS Catalog is enriched for the expression quantitative trait loci (eQTLs). The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation, including in the brain. Dr. Hancock mentioned multiple examples of genes associated with the stages of addiction and the specific regions of the brain in which they are expressed.


Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS
GTEx
eQTLs
GWAS
NIH National Institute on Drug Abuse
Advancing Addiction Science

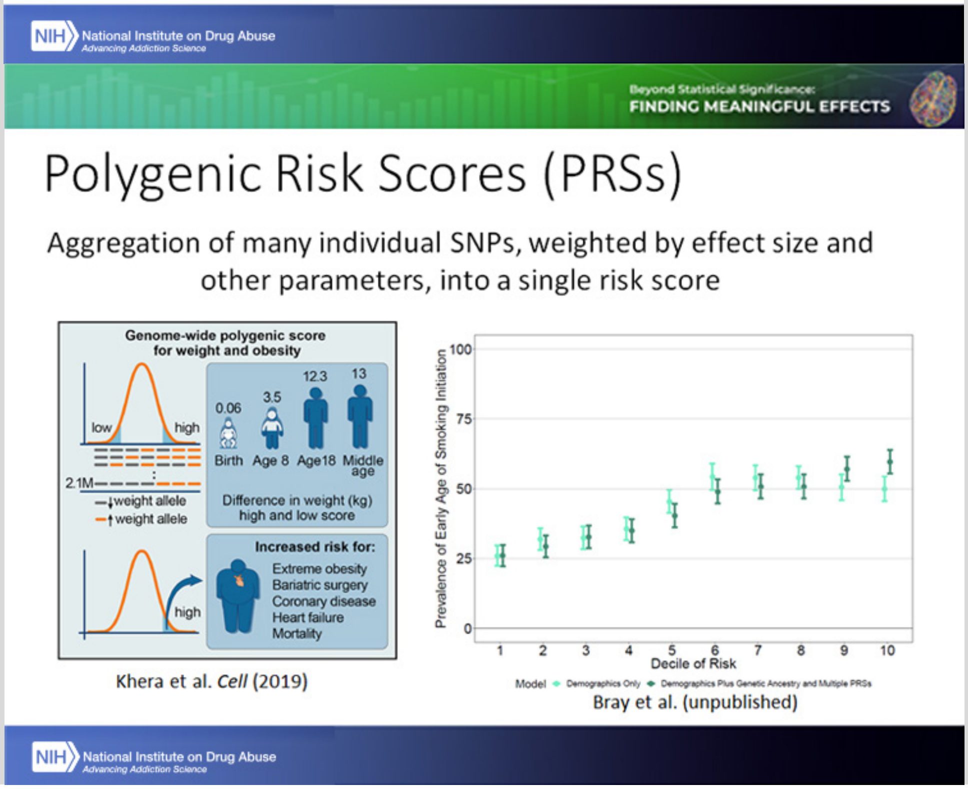
Regulatory Effects of Noncoding SNPs in Brain

Ever/never smoking GWAS (N=623k)
Liu et al. *Nat Genet* (2019)

S-PrediXcan method
Barbeira et al. *Nat Commun* (2018)

Tissue	Nucleus accumbens	Amygdala	Frontal cortex
Stage of Addiction Koob & Volkow *Lancet Psych* (2016)	Binge/ intoxication	Withdrawal/ negative affect	Preoccupation/ anticipation
Significantly associated genes	*C10orf32, RP11-18I14.10, TRPC6, SRR, TYW5, TMEM161B-AS1*	*CTC-498M16.4, AS3MT, C10orf32, RP11-18I14.10, SUOX, TYW5*	*C10orf32, TMEM161B-AS1, TYW5, AS3MT, SRR, CTC-498M16.4, ZKSCAN8*

with Luke Evans (unpublished)



Polygenic Risk Scores (PRSs)

Aggregation of many individual SNPs, weighted by effect size and other parameters, into a single risk score

Khera et al. *Cell* (2019)

Bray et al. (unpublished)

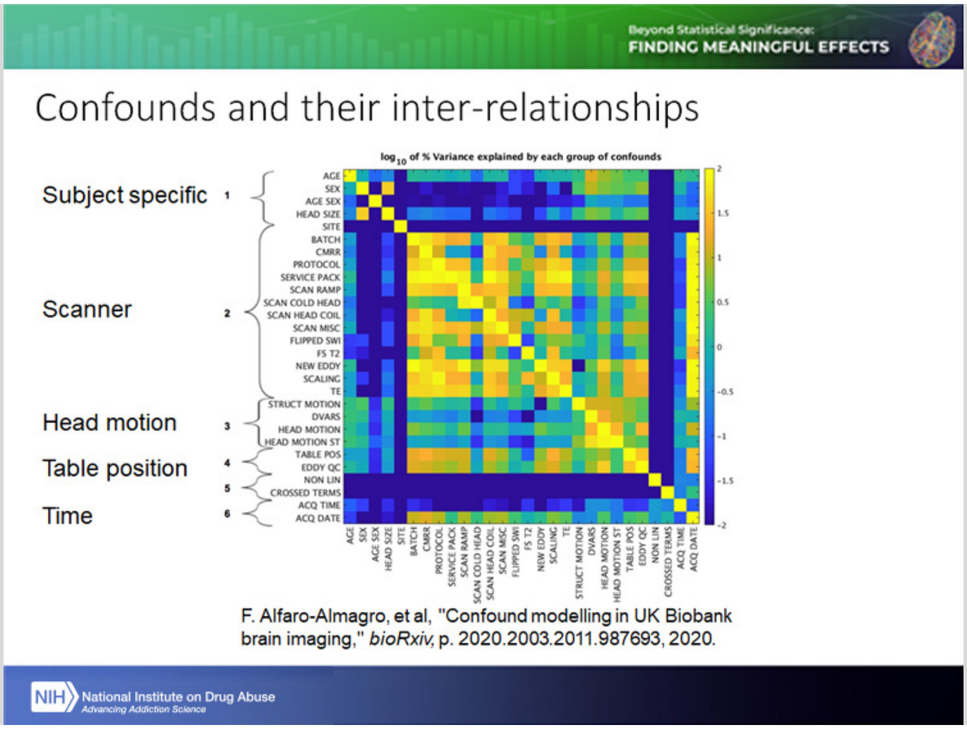
Clinically meaningful effects have also been found in the GWAS Catalog. Medication targets with GWAS support succeed twice as often as those without such support. For example, an intervention based on *CHRNA5* rs16969968 has been shown to enhance the effectiveness of smoking cessation. Polygenic Risk Scores (PRSs)—the aggregation of many individual SNPs, weighted by effect size and other parameters, into a single risk score—also have clinically meaningful effects (e.g., predicting obesity). In the addiction field, researchers have found many loci that are associated with smoking and alcohol consumption. In a value-added approach, researchers can layer PRSs over demographic characteristics for the purposes of prediction (e.g., younger age of smoking initiation). PRSs are advancing the field.

In summary, Dr. Hancock argued that GWAS-identified effects are statistically meaningful. The evidence for thousands of highly significant and replicable SNP-trait associations is irrefutable. These effects are also biologically meaningful, as GWAS reveals coding SNPs with functional effects on proteins. The vast majority of GWAS-identified SNPs are noncoding, with potential regulatory effects on target genes in relevant tissues. These effects are also clinically meaningful because approved medication targets are enriched with GWAS evidence. Additionally, SNP effects into PRSs can have predictive utility.

Covariates/Collinearity—Strategies to Model Confounds: Examples from Brain Imaging

Vince Calhoun, Ph.D.
Georgia State University, Georgia Institute of Technology, and Emory University

Dr. Vince Calhoun emphasized the need for researchers to define their terms. He noted a few key definitions. A nuisance variable, or confounder as correlated variable, is an unwanted variable that is typically correlated with the hypothesized independent variable within an experimental study, but is typically of no interest to the researcher. A control variable is an experimental element that is constant throughout the course of the investigation. A confounder as causal variable (lurking variable) is a variable that influences both the dependent variable and independent variable, causing a spurious association. Examples of types of confounds include cohort (e.g., age, sex, and diagnosis), timing (e.g., time of day, year, and ordering of scan), measurement (e.g., site, scanner, and rater), and function (e.g., motion, respiration, and eye blinks). Confounds can have complex interrelationships, which present challenges to researchers. Some less usual confounds that researchers often do not consider include

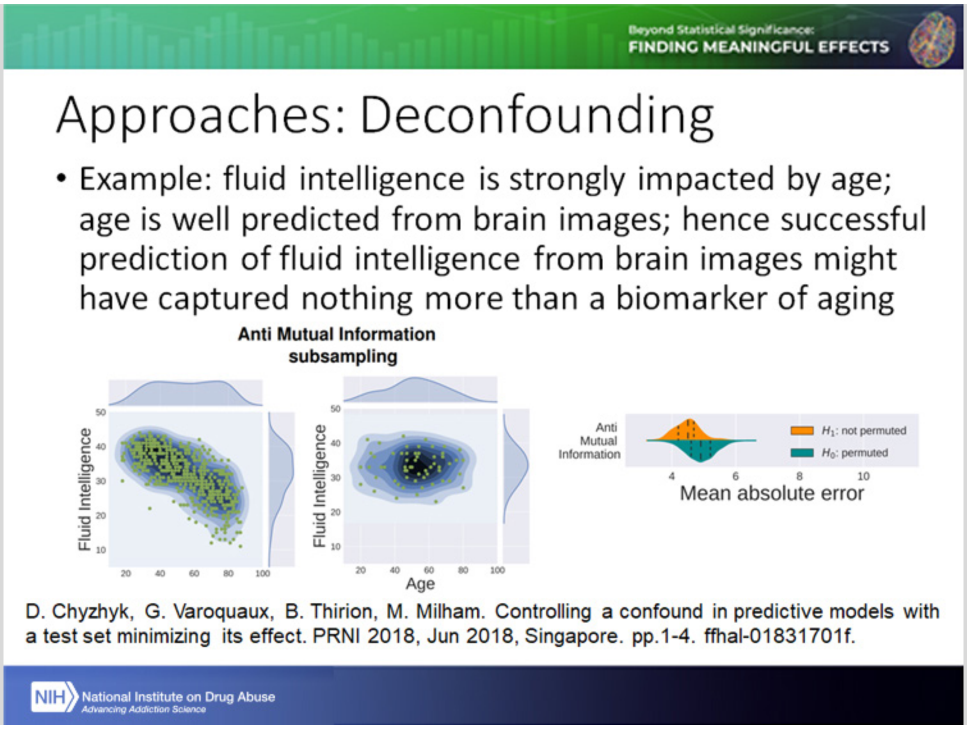


Confounds and their inter-relationships

F. Alfaro-Almagro, et al, "Confound modelling in UK Biobank brain imaging," *bioRxiv*, p. 2020.2003.2011.987693, 2020.

weather and seasonal variations. Confounds influence the brain estimates that are often used as independent variables. However, there are established confound mitigation strategies—such as preprocessing, calibration, harmonization across multiple sites, adjusting units, feature selection, visualization, impact analysis, and cross-validation and replication.

Researchers have various approaches and tools to address confounds. These include design elements, such as restriction, matching, and randomization. Analytical controls include regression, multivariate denoising, deconfounding, and deep learning—and these have increasing data requirements. Researchers can use regression/the General Linear Model (GLM) to adjust for confounds. Issues that arise are: (1) dependence between independent variables and control variables, and (2) dependence between control variables and an independent variable. In the GLM: Outcome = (effects of interest) + (confounds). *Is psychopathology associated with brain volume?* For this research question, the outcome is brain volume, and effect of interest is symptom severity. Confounding effects include age and gender. Dr. Calhoun emphasized that adjustment is not always the solution, as researchers may remove part of the effect of interest (e.g., adjusting for smoking while studying alcohol use). They must choose wisely when considering adjustment for confounds.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Approaches: Deconfounding

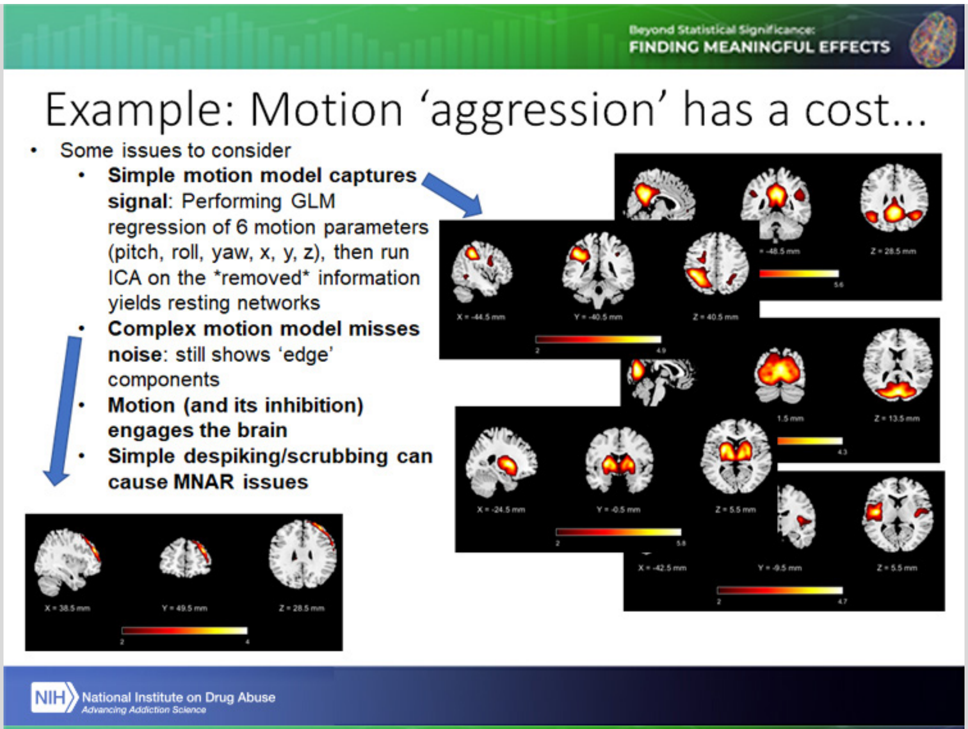
- Example: fluid intelligence is strongly impacted by age; age is well predicted from brain images; hence successful prediction of fluid intelligence from brain images might have captured nothing more than a biomarker of aging

D. Chyzhyk, G. Varoquaux, B. Thirion, M. Milham. Controlling a confound in predictive models with a test set minimizing its effect. PRNI 2018, Jun 2018, Singapore. pp.1-4. ffhal-01831701f.

NIH National Institute on Drug Abuse
Advancing Addiction Science

Considerations for assessing and modeling confounds include the measure (e.g., heart rate, age), estimate (e.g., motion), and variations (e.g., transformations, additive, multiplicative). Researchers must also consider whether confounds are characterized by linear or general dependence. GLM is for covariates with linear effects. Mutual information can capture nonlinear dependencies. For example, researchers can use multivariate independent component analysis (ICA) denoising to adjust for scanner effects. An approach to deconfounding involves deep learning with U-Net, a convolutional neural network, or generative adversarial networks, which are unsupervised learning algorithms. For example, fluid intelligence is strongly affected by age, which, in turn, is well predicted from brain images. Thus, successful prediction of fluid intelligence from brain images might have captured nothing more than a biomarker of aging. Deep learning can be used to address this problem.

However, confound adjustment has limitations. A particularly difficult confound to adjust for is motion during brain imaging. A simple motion model captures signal and researchers can perform a GLM regression of six motion parameters (i.e., pitch, roll, yaw, x, y, and z), then run ICA on the removed information, which yields resting networks. A complex motion model misses noise and still shows ‘edge’ components. Another issue is that motion (and its inhibition) engages the brain. Simple de-spiking/scrubbing can cause data missing not at random issues. Ideally, decisions should be automated. The GLM model on ICA functional network connectivity matrix allows researchers to visualize the impact of motion on resting fMRI at multiple spatial scales (e.g., 20, 50, 75, 100). Researchers can also harmonize large MRI data sets via a generalized additive model. They can use tools for interactive data visualization to view the impact of motion and other variables (scanner, site) on resting fMRI. Classification hybrid models adapt to individual subjects, preserve network labels, and address fuzzy parcellations.

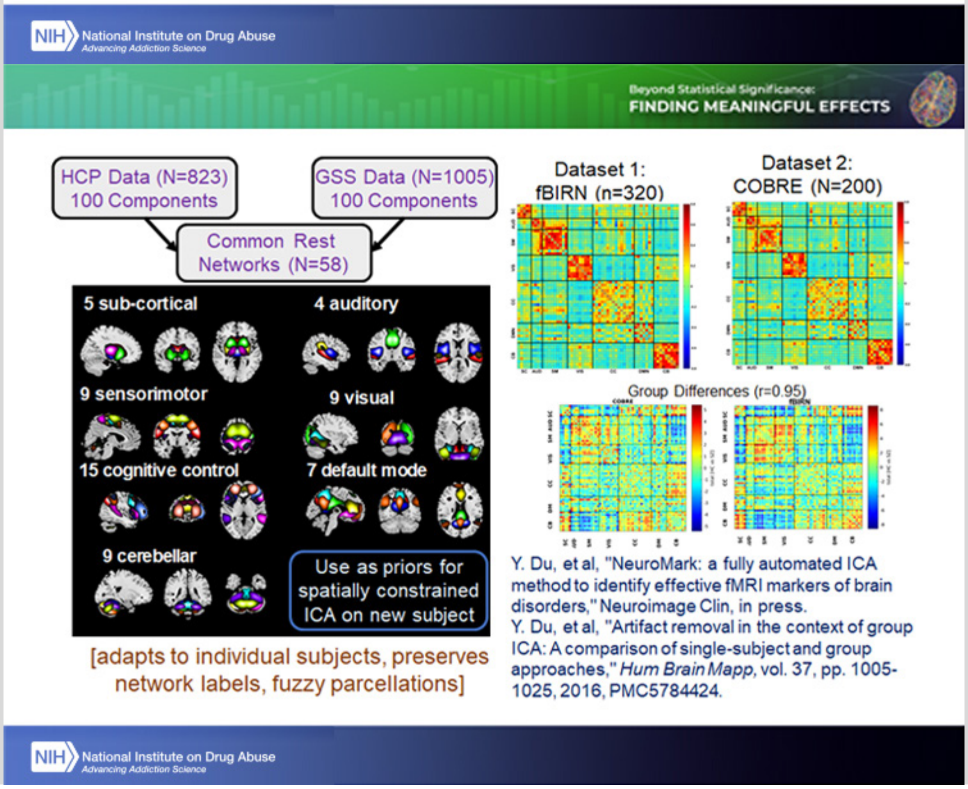


Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Example: Motion ‘aggression’ has a cost…

- Some issues to consider
  - **Simple motion model captures signal**: Performing GLM regression of 6 motion parameters (pitch, roll, yaw, x, y, z), then run ICA on the *removed* information yields resting networks
  - **Complex motion model misses noise**: still shows ‘edge’ components
  - **Motion (and its inhibition) engages the brain**
  - **Simple despiking/scrubbing can cause MNAR issues**

NIH National Institute on Drug Abuse
Advancing Addiction Science



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

HCP Data (N=823)
100 Components

GSS Data (N=1005)
100 Components

Common Rest Networks (N=58)

Use as priors for spatially constrained ICA on new subject

[adapts to individual subjects, preserves network labels, fuzzy parcellations]

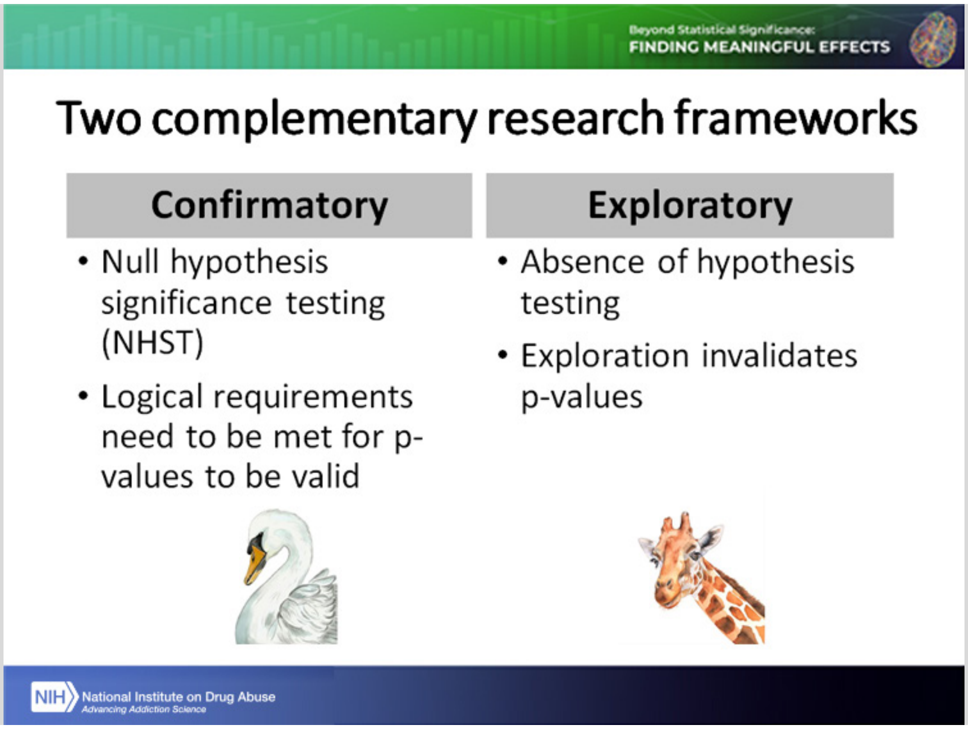
Y. Du, et al, "NeuroMark: a fully automated ICA method to identify effective fMRI markers of brain disorders," Neuroimage Clin, in press.
Y. Du, et al, "Artifact removal in the context of group ICA: A comparison of single-subject and group approaches," *Hum Brain Mapp*, vol. 37, pp. 1005-1025, 2016, PMC5784424.

NIH National Institute on Drug Abuse
Advancing Addiction Science

In summary, researchers should address confounds both during study design and analysis. There are many options for addressing/modeling confounds. Researchers should consider their options up front and be careful to avoid “garden of forking [confounds],” which can lead to bias. Big data offer the ability to incorporate and model larger sets of confounds. However, confounds can be averaged out, modeled, or amplified in studies with large sample sizes. There is no easy solution to dealing with confounds, but it is important that researchers visualize data and carefully consider the implications.

The Power of Boundaries: Confirmatory Versus Exploratory Research in Developmental and Clinical Neuroscience

Jennifer Pfeifer, Ph.D.
University of Oregon



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Two complementary research frameworks

Confirmatory	Exploratory
• Null hypothesis significance testing (NHST) • Logical requirements need to be met for p-values to be valid	• Absence of hypothesis testing • Exploration invalidates p-values

NIH National Institute on Drug Abuse
Advancing Addiction Science

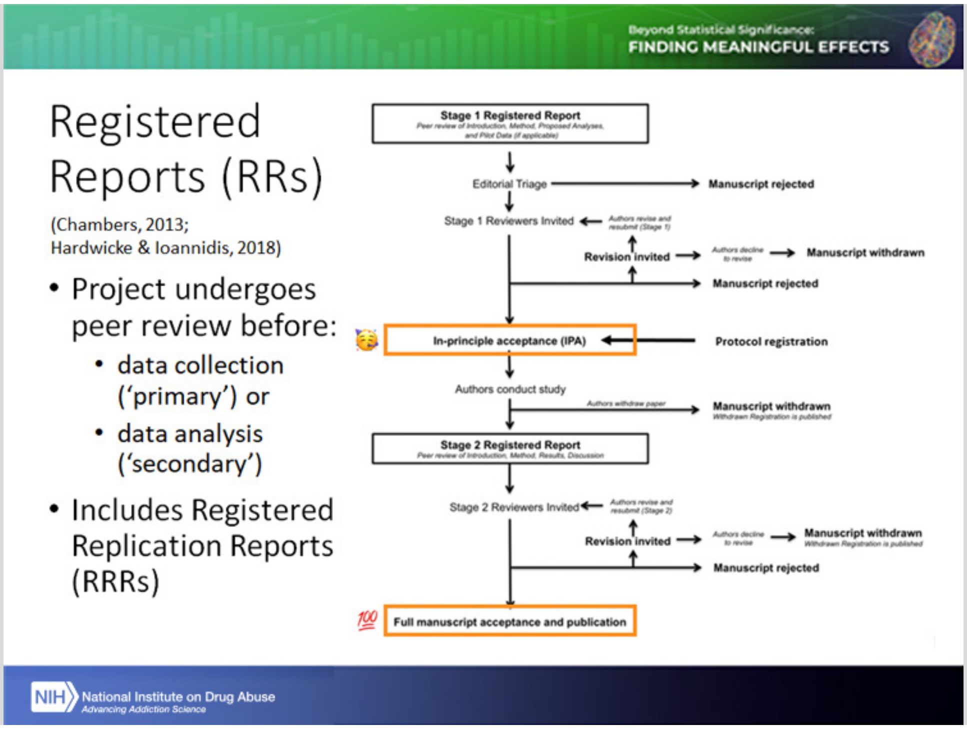
Confirmatory and exploratory work represent two complementary research frameworks. Researchers should understand the strengths and limitations of each as well as when and how to use both frameworks. Confirmatory research involves null hypothesis significance testing, and logical requirements need to be met for p-values to be valid. However, research sometimes does not meet these requirements. Historically, the field has depended too much on confirmatory techniques. In exploratory research, there is an absence of hypothesis testing, and exploration invalidates p-values. This framework can be valuable if built on a solid foundation. There is a need to correct for many researcher decisions. Researchers have some degrees of freedom and make analytical decisions after data have been observed. This results in a proliferation of analytical pathways, especially in big data studies.

> "Researcher degrees of freedom do not feel like degrees of freedom because, conditional on the data, each choice appears to be deterministic. But if we average over all possible data that could have occurred, we need to look at the entire garden of forking paths and recognize how each path can lead to statistical significance in its own way." (Gelman & Loken, 2013)

The "garden of forking paths" is an issue in neuroimaging. Researchers must make myriad decisions from the Digital Imaging and Communications in Medicine format for handling, storing, printing, and transmitting information in medical imaging to obtaining the results of studies. Each data-contingent decision increases the need for correction. If researchers do not know how many possible tests they could perform, the probability of false positives cannot be known or controlled. Investigating the effects of different parameters/pipelines typically results in inadequate correction for Type I error in traditional confirmatory methods. Researchers must also correct for many voxels. They must consider family-wise error (FWE) versus false discovery rate. Bonferroni correction at the voxel level and joint magnitude and spatial extent at the cluster level are two options for FWE correction. Importantly, cluster-correction techniques move inference and error control from voxel to cluster. One cannot make within-cluster inferences.

A limitation of p-values is that they may obscure effect sizes and distract researchers. In general, foundational cognitive neuroscience studies have been underpowered. Investigators should be mindful of the relationship between power and p-values. In big data studies, very small associations ($r< 0.05$) produce ‘compelling evidence’ to psychologists and cognitive neuroscientists for rejecting the null hypothesis ($p< 0.05$). A less obvious example is the difficulty in thresholding task-fMRI results from big data studies. Another limitation of p-values involves rejecting the null hypothesis. One can fail to reject the null even when there is a true effect. This can happen when the size of effect is too small to be detected given the methods (due to imprecise measurement) and the sample size. Researchers can only use a p-value to infer that there is some effect in certain voxels. Everywhere else there may be a difference that the study was too underpowered to detect. Retaining the null hypothesis requires an appeal to power, and here, big data will be useful.

Tools are available to protect researchers’ confirmatory analysis. Preregistration of manuscripts—specifying a research plan in advance of a study and submitting it to a registry—can be useful throughout the scientific process. Researchers can preregister their analyses at Open Science Framework at the Center for Open Science (https://cos.io/prereg/), which allows them to embargo and update their work over the duration of a project (time-stamped amendments). On the site, investigators can refer to their standard operating procedures and study protocols. Such a framework facilitates a more straightforward analysis, writing, and review process—helping to identify unknowns. The process of preregistration increases transparency and reproducibility.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Registered Reports (RRs)

(Chambers, 2013;
Hardwicke & Ioannidis, 2018)

- Project undergoes peer review before:
  - data collection ('primary') or
  - data analysis ('secondary')
- Includes Registered Replication Reports (RRRs)

Stage 1 Registered Report
Peer review of Introduction, Method, Proposed Analyses, and Pilot Data (if applicable)

Editorial Triage → Manuscript rejected

Stage 1 Reviewers Invited ← Authors revise and resubmit (Stage 1)

Revision invited → Authors decline to revise → Manuscript withdrawn

Manuscript rejected

🥳 In-principle acceptance (IPA) ← Protocol registration

Authors conduct study

Authors withdraw paper → Manuscript withdrawn
Withdrawn Registration is published

Stage 2 Registered Report
Peer review of Introduction, Method, Results, Discussion

Stage 2 Reviewers Invited ← Authors revise and resubmit (Stage 2)

Revision invited → Authors decline to revise → Manuscript withdrawn
Withdrawn Registration is published

Manuscript rejected

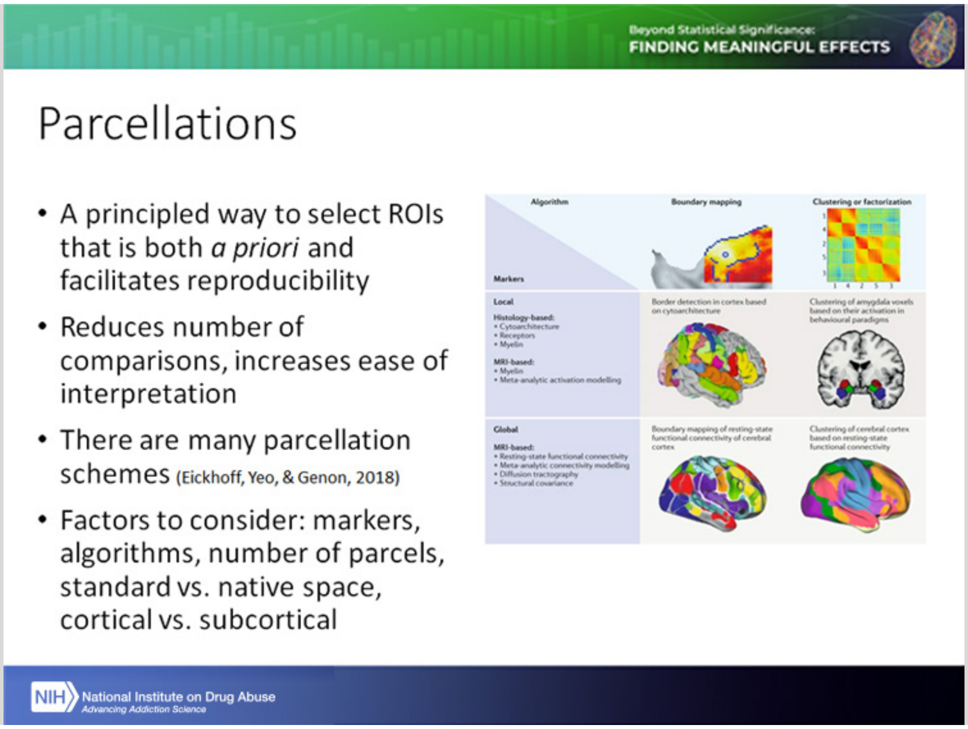
💯 Full manuscript acceptance and publication

NIH National Institute on Drug Abuse
Advancing Addiction Science

Investigators may also choose Registered Reports (RR), a publishing format that emphasizes the importance of the research question and the quality of methodology by conducting peer review prior to data collection. High-quality protocols may be provisionally accepted for publication if the authors follow through with the registered methodology. RRs may be for either primary or secondary data analyses and include Registered Replication Reports. Secondary RRs are submitted for peer review after data collection but before data analysis (e.g., ABCD study and Human Connectome Project). Secondary RRs are strongest when researchers provide evidence that the data have not been observed (e.g., access to data has been controlled). Knowledge of baseline data and cumulative changes from landmark large-scale studies will soon permeate the literature. An important question that the field must address: *How do we track what is known and weight its impact on analysis plans?*

The field needs enhanced specificity in its theories. The current practice of formulating theories in natural language using heuristic definitions of brain regions and their functions is ambiguous. Specifying a null-of-zero effect (“nil null”) allows researchers to accept even small significant effects as evidence for their hypothesis. Instantiating theories as computational models can complement or reduce need for preregistration.

In an exploratory data analysis framework, there is an absence of hypothesis testing. Exploratory work can range from model-free graphical visualizations to characterizations of ways in which fitted models depart from data. Open science tools make it easy to explore exhaustively one’s data (e.g., using R, BIDS apps). Researchers often focus on estimation and comprehensive reporting. Maps from group-level analyses are like any other set of variables, and tools provide summary statistics on them. In NeuroVault (https://neurovault.org), researchers can upload any statistic (t, F, beta, percent signal change) and get information on future power calculations and meta-analyses. Tools also provide a clearly scaled map of standardized effect sizes.



Beyond Statistical Significance:
FINDING MEANINGFUL EFFECTS

Parcellations

- A principled way to select ROIs that is both *a priori* and facilitates reproducibility
- Reduces number of comparisons, increases ease of interpretation
- There are many parcellation schemes (Eickhoff, Yeo, & Genon, 2018)
- Factors to consider: markers, algorithms, number of parcels, standard vs. native space, cortical vs. subcortical

Markers / Algorithm	Boundary mapping	Clustering or factorization
Local **Histology-based:** • Cytoarchitecture • Receptors • Myelin **MRI-based:** • Myelin • Meta-analytic activation modelling	Border detection in cortex based on cytoarchitecture	Clustering of amygdala voxels based on their activation in behavioural paradigms
Global **MRI-based:** • Resting-state functional connectivity • Meta-analytic connectivity modelling • Diffusion tractography • Structural covariance	Boundary mapping of resting-state functional connectivity of cerebral cortex	Clustering of cerebral cortex based on resting-state functional connectivity

NIH National Institute on Drug Abuse
Advancing Addiction Science

Brain parcellations are a principled way to select regions of interest because they are both *a priori* and facilitate reproducibility. This method reduces the number of comparisons and increases the ease of interpretation. There are many parcellation schemes and in choosing, researchers must consider markers, algorithms, number of parcels, standard versus native space, and cortical versus subcortical regions. Parcellations can be applied to data at different levels (i.e., group, individual, and trial). But using parcellations has some important limitations, including interindividual variability and that they have usually been created from adult data.

Specification curve analysis (SCA), which is also called multiverse analysis, addresses the problem of researcher degrees of freedom—that is, that there are many ways to test the same research question. In big data studies with many variables and researchers, this problem snowballs. SCA allows researchers to quantify and visualize the stability of observed effects across many possible models, offering a kind of sensitivity analysis.

In conclusion, Dr. Pfeifer argued that the field needs both exploratory and confirmatory frameworks. The rigor of confirmatory analyses should be bolstered and more value should be accorded to exploratory analyses. She encouraged participants to use preregistration and RRs to constrain researcher degrees of freedom and increase transparency and reproducibility. The merits of exploratory work are vast, and researchers should learn how to conduct and review these analyses.

Breakout Sessions

Discussants: *Dr. Todd Constable, Dr. Erin Dunn, Dr. Amit Etkin, Dr. Damien Fair, Dr. Robert Grossman, Dr. Mike Hawrylycz, Dr. Andrew Gelman, Dr. Kevin King, Dr. Ben Lahey, Dr. Beatriz Luna, Dr. Gina Lovasi, Dr. Kate Mills, Dr. Max Owens, Dr. Clare Palmer, Dr. Bogdan Pasaniuc, Dr. Martin Paulus, Dr. Joshua Sampson, Dr. Wesley Thompson, Dr. Lucina Uddin, Dr. Lorenzo Vega, Dr. Ragini Verma, Dr. Ashley Watts*

All discussants and participants rotated through breakout sessions on three topics: (1) small effects, (2) covariates/collinearity, and (3) exploratory and confirmatory data analysis frameworks. NIH staff members facilitated breakout sessions, and panelists and discussants highlighted key points in their respective areas of expertise. Participants submitted questions and comments.

Charge & Logistics

Dr. Hoffman explained that the overall objective of the breakout groups was to synthesize best practice approaches for identifying and analyzing meaningful effects as they apply to biological and medical research. She reviewed the questions for each of the breakout sessions, which were intended to guide but not limit discussion.

Summary Breakout Reports

Small Effects

Dr. Dana Hancock

Dr. Hancock reviewed the overarching questions considered in the breakout session.

How can small effect sizes be interpreted in terms of causality or prediction?

Regarding neuroimaging and GWAS effect sizes, breakout group participants distinguished between biology and “clinical” prediction, and emphasized the need to state explicitly the research question. Researchers use GWAS to guide hypothesis-driven research and then use multiple lines of evidence to prioritize the results (even with small effect sizes) for experimental studies. Additional comments included:

- It is crucial to understand that effect size is not equivalent to being able to draw a causal inference.
- Large effect sizes are not always “meaningful.”
- A question is whether small effect sizes can be causal, and this requires a change in researcher mindset.
- Including covariates (e.g., comorbidities) can make small effects even smaller.
- The field is not yet in a position to conduct thresholding for effect sizes.
- Phenotype measurement in observational studies affects effect sizes.
- Data from electronic health records offers opportunities, but can lead to different conclusions.
- Gene-environment or other higher-order interactions attenuate effect sizes.
- One must consider the representativeness of exposure and the differences in observational and experimental research settings.

Should there be different standards when interpreting results in terms of a detectable effect versus an effect that could be the basis of an intervention?

This question depends on whether the research involves clinical prediction and/or intervention. Here, the effect sizes within subgroups is important. Effects are sometimes context dependent (e.g., environmental exposure).

When is a nonlinear analysis justified in evaluating a small linear effect?

There are limitations of a linear model. For example, do researchers need to fully explain variance in outcomes? An issue is non-normal distributions (zero inflation). With big data and gene networks, there is a need for systems modeling. In longitudinal analyses, the impact of linearity may be particularly important on time and developmental period.

Covariates/Collinearity

Dr. Vince Calhoun

Dr. Calhoun reviewed the guiding questions discussed by the breakout group participants.

Determining which covariates to include in statistical models is complex and nuanced, especially in large data sets. Given the impact of covariate selection on replicability and reproducibility, these decisions must be made thoughtfully. Are there optimal strategies for selecting covariate controls? What factors must be considered?

The breakout group participants determined that there is no easy answer to selecting covariates. However, researchers can think carefully about their research questions and how to address covariates. Covariate selection is highly dependent on the model being tested. The use of domain knowledge is critical in this process. Depending on the research question, a variable might or might not be appropriate to include as a covariate. A good approach is to divide possible confounds into batches—essential covariates that are needed for interpretation, possible confounders that may be mediators, and more speculative effects.

Considerations include that site effects are not limited to scanner effects. Relevant site effects include demographics and recruitment differences. Often, a variable is a proxy for many other factors. Reporting bivariate relationships among confounds is good. Participants in the breakout session have found that even apparently simple effects can be complex. For example, age influences motion in the scanner, which is also related to intelligence, a function of age.

To evaluate covariates, researchers can run an analysis with and without covariates (sensitivity analysis) and examine whether the main effects change dramatically. This can also help determine whether a variable should be considered a confound or a mediator. Researchers can also use subsets of data to evaluate possible confounds and refine their model, then apply the model on the hold-out data. Cross-validation/replication of results also evaluates covariates.

Participants in the breakout session generally rejected the notion of requiring default covariates. Some covariates could be recommended, but not required. The ABCD study's matching variables (site, sex, marital status, etc.) are a possible model. However, these may not be appropriate depending on the question of interest. The ABCD model will likely not generalize to other studies. A given study might be able to develop a recommended list of possible confounds.

To minimize the impact of covariates, researchers can use machine learning approaches to eliminate confounds in the data. Researchers can also evaluate the size of an effect that would be needed for an unobserved confound to eliminate the effect of interest. A large sample size enables researchers to have more precise estimates. Effects tend to stabilize in sample sizes of 1,000 to 2,000. Researchers should keep in mind that variation can be random or systematic. Systematic variance cannot be fixed by sample size.

Regarding reporting results, participants agreed that it is important for researchers to be transparent about all analyses conducted—including non-significant results. Researchers should release the code to allow others to replicate their results. They should also include the results of such analyses in supplemental materials. References discussed in the breakout session include https://www.acpjournals.org/doi/10.7326/M16-2607 and https://arxiv.org/abs/1805.06826.

Exploratory vs. Confirmatory Data Analysis Frameworks

Dr. Jennifer Pfeifer

What are perceived obstacles to submitting registered reports and what can be done to encourage it?

Preregistration and registered reports create rigor and transparency (see Open Science Framework – https://osf.io/). These frameworks show the value of multidisciplinary collaboration and team science that can balance domain knowledge and data science, and they can protect researchers against publication bias for non-null findings. However, there are many misperceptions in this area. The situation is complex, as exploratory analyses inform confirmatory approaches, and the boundaries are fluid.

MYTH	REALITY
Preregistration and registered reports constrain intellectual freedom	**Exploration is allowed in preregistration and registered reports**
Preregistration and registered reports are rigid, and researchers can't update best practices	**Preregistrations and registered reports can be amended and do not preclude changing methodological directions or correcting mistakes**

Overall, the process requires a culture change:

- Individuals and teams need to change their mindsets
    - Individuals need mentorship and training in these frameworks and other open-science practices
    - Collaborations may be retrospective which may increase evidence for convergent or divergent validity
- Journals
    - Journals need to recognize the value of preregistered or registered reports to understand that while they are more likely to produce null results or small effects, these approaches typically lead to high citation rates (see https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2019.01299.x and https://docs.google.com/spreadsheets/d/1gDk6bQLT9fvH-J67uPQpRnh250eK2ZocyfGzZKlmkK4/edit#gid=0)
    - Editors and reviewers need to move away from asking authors to hypothesize after the results are already known or add confirmatory hypotheses in exploratory contexts
- Research Institutions
    - Allocating "credit" for the work involved is complex, particularly in promotion and tenure systems
- NIH and other funding organizations
    - Encourage applications using exploratory approaches
    - Encourage applications using holdout samples to address the issue of overfitting data in exploratory analyses
    - Create data science resources—such as centralized practices and analytics and standardized methods for data outputs

Community-wide acceptance of these frameworks is essential, and their adoption should not be primarily driven by early career researchers. Ethical science is a shared responsibility across levels—including individuals, teams, editors, reviewers, publishers, and funders.

Grand Discussion

Is it possible to advance effect-size thresholding?

Dr. Hancock commented that the field has not made enough progress on this approach to adopt it. A better understanding of biology would permit the overlaying of effect sizes for clinical utility to improve decision making, but the evidence base is far from that point. Dr. Calhoun added that for imaging data, it is useful to have translucent brain maps (with incomplete thresholds) to see patterns in regions that are below the threshold. A machine learning approach can assist with decoupling related confounds. Researchers can try to select subsets of samples from a larger data set, decorrelate variables, and then examine variables of interest.

Are researchers using patient input outcomes and experiences to inform research?

Dr. Holly Moore remarked that qualitative research with patients can provide important information.

> ...exploration is a significant advantage of large databases. After exploratory analyses, researchers can conduct a deep analysis that if replicated, points to a non-random effect. (Dr. Nora Volkow)

What is the appropriate time to cease exploratory analysis?

Dr. Pfeifer focused on *a priori* outlining the model as specifically as possible followed by an analysis to identify the problem space. Developing a descriptive a priori plan for starting and stopping analyses fosters comprehensive transparency. Dr. Calhoun commented that there are degrees of exploratory analysis. One might have a broad hypothesis (e.g., whole brain effects) or more constrained hypotheses on a small number of variables. Dr. Volkow remarked that exploration is a significant advantage of large databases. After exploratory analyses, researchers can conduct a deep analysis that if replicated, points to a non-random effect. Replication is crucial in determining the validity of observed effects. Although this approach differs from traditional science, she encouraged researchers to be bold in addressing the complexity of variables.

Participants discussed the value of the preregistration of analyses, including for replication and determining the robustness of findings. Dr. Damien Fair commented that one lesson of exploratory analysis is that tests that yield significant effects often are not able to be replicated in small samples, but may be in large samples. That is, effect sizes can be inflated in exploratory analyses with small samples. Dr. Calhoun added that early imaging studies yielded small trait-based effects, and researchers need large sample sizes to test these results due to high variation and the conflation of variables. Researchers have not yet fully determined the robustness of these findings. The ABCD study has released pre-split samples for exploratory analyses followed by replication. Dr. Calhoun pointed out that replicating within a harmonized study is not a complete replication. Additionally, an attendee noted that there are few projects with sample sizes as large as the ABCD study.

What is the best way for NIDA to invest resources and balance studies with large data sets and smaller independent projects?

Dr. Calhoun responded that both approaches are needed—determining robust effects in large studies, then replicating findings in small independent projects. Dr. Fair agreed that both types of studies are important, adding that NIDA must be strategic about the questions asked in small independent studies, which should be coordinated so that they are amenable to subsequent combination. Dr. Volkow commented that the ABCD study allows other researchers to build off the findings related to prevention and intervention to advance knowledge. The ABCD study's large data set is sensitive to the effects of socioeconomic status and health disparities, offering opportunities for replication with smaller independent cohorts. Dr. Fair commented that small effect sizes can be caused by population variability or differences in measurement (how the data are collected). This issue is complicated, but researchers can use within-subjects designs to test interventions and increase statistical power.

> Participants recommended including underrepresented groups in large population studies to the greatest extent possible with the goal of developing models that predict for everyone (not for the purposes of comparison).

An attendee commented on the pressing need for studies with diverse populations. As study populations include mostly people of European heritage, the generalizability of the findings to other groups is a concern. Dr. Volkow remarked that for the ABCD study, power is lacking for some groups despite oversampling. The population distribution for the study is related to the location of its research centers, and there are more participants with higher socioeconomic status, which presents challenges. The Helping to End Addiction Long-term^{SM} (HEAL) Initiative's HEALthy Brain and Child Development (HBCD) study—which will establish a large cohort of pregnant women from regions of the country significantly affected by the opioid crisis and follow them and their children for at least 10 years—is working to overcome these issues. An attendee noted that some genetics projects are working to genotype diverse populations, but it will take time for the data to be collected and made available for analysis (e.g., GWAS). Participants recommended including underrepresented groups in large population studies to the greatest extent possible with the goal of developing models that predict for everyone (not for the purposes of comparison). Participants also discussed the great influence of the environment on the brain, emphasized the importance of phenotypic assessment, and noted the need for data-driven brain parcellation atlases.

Dr. Lucina Uddin remarked on the need to educate grant reviewers about preregistration of analyses and registered reports (e.g., that this science takes more time). Dr. Hoffman agreed that this is an important topic and NIH needs feedback from researchers on suggested approaches to change the culture. Dr. Volkow suggested that NIDA create a chat box on its website to obtain input from the field on the best ways to allow for flexibility and creativity in big data. Researchers could also send feedback to their NIDA Program Officers.

Participants discussed approaches to small studies. Dr. Grant remarked that researchers could concentrate on variables that seem to drive the observed effects. Dr. Wesley Thompson focused on the potential of research design to increase effect size, as mechanism is not the only driver. He suggested leveraging large studies to provide data sets for smaller ones, as in research on polygenic risk. Dr. Beatriz Luna remarked that small studies are valuable because they explore innovations, but the question is how to interpret the results. Dr. Ben Lahey commented that the effects sizes in studies on cognition are small, as would be expected, because many variables account for variance. Dr. Clare Palmer noted that small studies must have sensitivity and power so that researchers can be confident that the results are not spurious. The ABCD study can estimate the variance structure, then researchers conducting a smaller study can boost power by taking that estimate into account. An attendee asked: *When causal manipulations are spatially and temporally precise, does one need a large number of subjects?* Dr. Thompson responded that sample size is separate from causes. The sample size needed depends on the effect size. In a natural setting, most effect sizes will probably be small.

Should researchers revisit Cohen's D?

Dr. Thompson stated that he has been considering Cohen's d, and at issue is what constitutes a practical meaningful effect. Most ABCD study researchers are focused on scientific discovery rather than clinical applications. Effects sizes are small for many reasons, and researchers should not ignore every finding with a small effect. Dr. Pfeifer suggested training peer reviewers (of scientific articles and grants) and journal editors in best practices regarding power and effect sizes so that they can develop the expertise to evaluate whether research can answer the questions posed. Dr. Kevin King focused on the need for methods research and engaging researchers in best practices. He suggested broad incentives for methodological research. An attendee encouraged participants to volunteer to serve as grant reviewers, as those with statistical expertise are always needed in this role.

> ...a prospective and longitudinal study, ABCD offers information on the baseline characteristics of subjects. Researchers can develop and test predictive models on the relationship between particular variables and a specific brain or behavioral trajectory. This is a powerful approach that adds complexity to the study.
> (Dr. Nora Volkow)

Dr. Volkow commented that as a prospective and longitudinal study, ABCD offers information on the baseline characteristics of subjects. Researchers can develop and test predictive models on the relationship between particular variables and a specific brain or behavioral trajectory. This is a powerful approach that adds complexity to the study. Dr. Thompson remarked on the need to examine within-person effects over time once those data are collected. This should be a focus of the ABCD study (and smaller studies) in the future.

Can researchers use propensity scores to get a sense of causality?

Dr. Thompson noted that there is always skepticism about causality in observational studies (every exposure is not randomly assigned and, thus, could be due to covariates). One can control for confounding variables, but there are an infinite number of covariates for which researchers cannot account. Dr. Lahey agreed and commented that the use of representative samples and quasi-experimental designs can provide the opportunity for causal inferences to some extent. Several attendees stressed that multiple lines of evidence are crucial for increasing confidence in a finding.

Wrap-Up

Dr. Volkow thanked Dr. Hoffman and all meeting organizers and participants for their thoughtful comments. The field is just beginning to optimize making data and resources available for analysis. It is important for NIH to facilitate access to training on analysis and methods to take advantage of the large ABCD study (and eventually the HBCD study) databases to advance the science on brain development. Dr. Hoffman also thanked participants and emphasized the meeting was a first step in the discussion on best practices and challenges.

Appendix: Detailed Breakout Session Notes

Small Effects

Panelist: *Dr. Dana Hancock*
Concept presenter: *Dr. Mike Hawrylycz*
Facilitators: *Dr. Elizabeth Hoffman, Dr. Steve Grant, Dr. Melissa Rotunno*

How should small effect sizes be interpreted in terms of causality or prediction? For example, does a small effect size in an observational study necessarily mean that a subsequent experimental manipulation or intervention will not be effective, or could not serve as an accurate outcome predictor?

To start the discussion, Dr. Grant posed several questions mentioned by Dr. Erin Dunn: *Should researchers account for 100 percent of variability in a study? Is that reasonable? Are researchers only interested in strong effects? Under what conditions would a very small effect be useful to understand experimental manipulation?*

Dr. Hawrylycz commented that many neuroscientists find that most cell types in the brain are rare. However, coherent patterns emerge after profiling cells and their genomes, clustering the data, and comparing with similar profiling in epigenetic data. Researchers do not yet understand these patterns or have a quantitative estimate of cell types. The question is: *What should researchers make of small effects under these conditions?* Small effects are important but require special treatments, including multimodal reproducibility, but nothing prevents comparatively small effects from being biologically meaningful.

Dr. Grant noted that this topic dovetails nicely with Dr. Hancock's points about different levels of meaningfulness. The criteria for statistical meaningfulness (e.g., p-values and power) are well established, but those for biological meaningfulness are unclear. Reproducibility seems a reasonable criterion. Criteria for clinically meaningful effects—either for individual patient improvement or population benefit—may be somewhat better established. However, criteria for the middle part of our understanding—that is, biologically meaningful effects—are currently unclear.

Dr. Hancock remarked that the effects from GWAS from simple regression models do not account for the likely impacts of much larger interactions or networks. As Dr. Volkow noted in the main session, scientists need a better understanding of those networks. The small effect sizes observed are partially due to not addressing networks.

Dr. Thompson pointed out that GWAS explains approximately half of variability. Studies with large sample sizes often find small effects that are replicable. In those cases, the question is: *How can these work together to suggest a phenotype?* This is a difficult but necessary area to address. It seems that imaging research has not progressed as far as GWAS in this respect. Perhaps imaging researchers should focus on identifying which parts of the brain are more strongly associated with particular behaviors by using GWAS as a model.

Dr. Hancock remarked on another complexity with imaging (and epigenetics). These studies simply measure at a single point in time. Therefore, effect sizes differ for the developing brain compared with the adult brain. Epigenetic studies also have small effect sizes, as do those with single cells.

Dr. Amit Etkin commented that, for small data sets, a concern is the level of characterization for phenotypes. GWAS becomes less meaningful as phenotype becomes more diffuse, and small effects may be hiding bigger effects.

> ...issues such as test-retest reliability and developmental versus cross-sectional designs need to be at the forefront of discussions. (Dr. Wesley Thompson)

Dr. Thompson added that measurement is a crucial aspect of why effects are small—issues such as test-retest reliability and developmental versus cross-sectional designs need to be at the forefront of discussions.

Dr. Lahey remarked that causality is an issue of research design. Observational versus experimental research design plays a significant role in assigning causality, and he asked the panel: *When the effect size is small, is it legitimate to call something a cause?* In his view, these are two related and unrelated concepts. One can have a very large effect size with no basis for causation, but also have a very small effect size detectable in research designs that permit causal inferences. Essentially, one cannot infer causation from correlation. Dr. Thompson noted that the clearest example of this is GWAS, for which all effects are small, but presumably effects in SNPs affect phenotype. Dr. Lahey commented that there is still some doubt about the pathways through which SNPs exert causal effects. A speaker for the panel session mentioned being careful about not controlling for environmental effects that are heritable. Therefore, researchers should be careful of gene-environment correlation. For example, one might find a set of SNPs that increases the likelihood that an individual will be physically assaulted. In this case, the causal pathway may not be from SNP to depression, but from SNP to environment to depression. Magnitude of effect size has nothing to do with whether you can draw a causal inference.

Dr. Grant summarized the discussion as: If a researcher observes a large effect size, it would be reasonable to follow up with an experiment to determine causality. Whereas, if the result has a very small effect size that accounts for a fraction of percent of variance, would one expect a larger effect size in a subsequent experimental study? Would manipulation of a variable that accounts for a small amount of variance be reasonably expected to point to causality? Dr. Lahey commented that large effect sizes are easier and offer more design options, but one could design an intervention to see whether reducing particulate matter in the atmosphere reduces ADHD based on a small effect size. Such a study would require a large sample, but it might be beneficial at the population level.

Dr. Grant focused on the question: In a study with a large sample, a researcher observes that manipulation predicts a very small change. *Is this causal?* An attendee stated that it depends on the design of the experiment. Dr. Thompson mentioned the concern that small effects are more likely to be confounded. Additionally, controlling for confounds may eliminate small effects. Dr. Grant remarked that the problem is not determining the causal nature of a factor that has an appreciable effect. Dr. Hawrylycz asked: *Are researchers going to solve these problems by only considering large effect sizes?* Neurobiology is sufficiently subtle, so scientists cannot avoid the problem of small effects. The best course of action is to create guidelines for prioritization and an approach for these small effects.

Dr. Thompson asked Dr. Hancock: *What can neuroimaging learn from prioritization of SNPs?* Dr. Hancock emphasized the importance of multiple other layers of data to inform SNP findings (e.g., as has been done with nicotine dependency and CHRNA5 [Cholinergic Receptor Nicotinic Alpha 5 Subunit]). Prioritizing SNPs requires not only evidence beyond GWAS statistical signals, but also the understanding that multiple independent SNPs can change. The question researchers face is which SNPs to prioritize (e.g., which to take to a mouse model). Answering this question requires layering on multiple lines of evidence. Information about pathways is based on this approach. But there is much more to be discovered with machine learning and statistical methods, both within the data and by layering out multiple omics and patterns. Dr. Jonathan Pritchard's group has published an article on building those networks and within them core genes and peripheral genes [https://pubmed.ncbi.nlm.nih.gov/28622505/]. This approach is useful for developing drug targets. There may be 100 genes involved in a network, but targeting a core gene can overcome its small effect size by layering multiple lines of evidence.

Dr. Thompson commented that an intervention could have a stronger effect than observation. Manipulating a variable can have a large effect even if measurement disguises effect size. One cannot assume that a small effect size means the causal effect of a variable is small.

Dr. Etkin stressed that inferring causality from observation is limited regardless. Additionally, the effects of interventions are not always aligned with discovering clear lines of causality. Researchers are approximating an outcome—how the intervention affects parts of the brain and body, in which way, the timing, and impact on a particular target. The interactions and complexity are massive. One intervention study is likely not sufficient evidence of a true causal association, as many factors are relevant (e.g., dosing). Rather than looking for "answers," researchers should determine goals related to conclusions that can be drawn from studies (and then the next steps) instead of worrying about clear associations.

An attendee asked: *Does the big-data approach no longer try to approximate population effects?* Dr. Thompson explained that, no, the goal of machine learning is to estimate something that replicates outside of your sample. Dr. Grant remarked that big-data approaches explicitly estimate population effects and yield more precise estimates of those effects. Dr. Hawrylycz emphasized the accuracy of an earlier comment that big data does not replace hypothesis-driven science. His impression of unstructured big data is that it produces better confidence intervals and predictability for very large effects and results that cannot be leveraged. Therefore, to have a clinically meaningful effect, one has to "place a bet."—that is, collect data subject to a hypothesis, because the exploration space is far too large. Big data will not solve that problem. Researchers need structured experiments and a great deal of data and models.

Dr. Grant asked whether it was correct that big data includes many variables on individual subjects, whereas population data features a very large sample size (many individual subjects). Dr. Hawrylycz noted that those terms are often confounded. "Big data" is often used as a proxy for detailed experimental design in which many variables are measured and researchers examine the data for various subtle effects. Even data collected from a large population requires a hypothesis. Both big data and population data have a role in research. However, by collecting more variables, the space to search for signals is large, and space for effects is multidimensional. Researchers who do not use domain knowledge and prior data end up with tighter and tighter bounds. Dr. Thompson added that it becomes harder to interpret observations with small effects in that situation. Dr. Lahey advocated for testing hypotheses in large data sets.

Dr. Grant asked whether characterizing 100 percent of variance in big data is a detriment, as one has to overfit the model and cannot generalize because of idiosyncrasies of the sample. Dr. Thompson remarked that doing so would require replication outside of that sample. He added that one could never account for 100 percent of the variance.

Dr. Grant gauged consensus among participants on what effect size is needed to consider it biologically meaningful. Dr. Hancock remarked that across the range, truly functional effects in addiction science vary from very small to large. One can find examples across that spectrum. Dr. Hoffman stressed the importance of considering whether the effect size is from one study or has been observed in multiple studies. Dr. Grant commented that in an observational study, one cannot a priori eliminate a causal role of a variable due to small effect size. But he asked: *At what point does one consider an effect trivial and not worth following up?* Participants agreed that this topic is important and requires further research. Dr. Thompson emphasized that effect size alone is not enough to make a judgement. Rather, effect size must be placed into the context of both theory and the overall body of evidence. Many factors contribute to a small effect size (measurement). Dr. Lahey added that small effect sizes are the norm, as researchers measure a huge number of variables and are not accounting for interactions. They should not be disappointed by small effect sizes.

Should there be different standards when interpreting results in terms of a detectable effect with predictive utility, but no intervention potential versus an effect that could be used as the basis for an intervention?

Dr. Grant commented that standards for a detectable effect might include whether it is reliable, replicable, and statistically acceptable. The larger issue is the interpretation of small effects and whether researchers consider a small effect to have practical utility. Should researchers be explicit and state that the effect would have practical utility and, therefore, evaluate it according to a standard or say that the observation contributes to a growing body of knowledge?

An attendee asked: *Why is there an assumption that the relevant effect size relates to clinical prediction?* Clinical prediction is only one of many uses of data. At this stage, the prediction of a substance use or psychiatric disorder carries potential stigma and discrimination. Therefore, the standards for deploying prediction should be extremely high. The field is nowhere near that point currently, despite the fact that some such proposals have been made. At this stage, the primary utility of GWAS and related approaches is to point robustly toward the relevant biology, which can then be explored. Such exploration can lead to clinically meaningful results other than prediction.

Dr. Todd Constable noted Dr. Gordon's point made in introductory comments about the distinction between clinical application and learning about the underlying biology of conditions. Small effects reveal information about biology, and combining multiple small effects can contribute to a model that produces a large effect. Small effects should not be ignored.

An attendee remarked that PRSs cover very small variations, but if researchers are pointed to the right pathway by them, the impact of a drug or other manipulation can be significant. Tiny effects can generate the perception that prediction is possible. Dr. Grant asked about distinguishing overhyped results from those that advance our understanding of biology given the same effect size. An attendee commented that in the context of prediction, GWAS discoveries are very tiny effects of common variance because of evolution. Some have used PRSs to distinguish cases from controls as opposed to clinical methods. However, these comparisons revealed no measurable differences, which suggests that the field is not ready to use PRSs for prediction.

An attendee asked for advice on solving problems with estimating effect sizes in neuroimaging. Dr. Grant stated that ABCD researchers are working on this analysis issue. Their approach will allow researchers to develop brain imaging maps that show effect sizes and variance. An attendee asked whether it is reasonable to assume that the vast majority of effects are small when considering the entire population and ecologically valid studies. Perhaps researchers are concerned about small effects because big data is a relatively new phenomenon. Dr. Martin Paulus suggested that there are two relevant issues. The first issue relates to small effect sizes with larger samples. There is some evidence that as sample size increases, brain-behavior relationships measured by correlation coefficients get smaller. This is seen when splitting ABCD samples into smaller groups. Second, there is an assumption that researchers are reporting false positives, driven by publication bias. This may not be true, and researchers need to separate inherently small effect sizes from large effect sizes that are false positives and large effect sizes that could be true. Dr. Thompson has published on this issue. In regression/correlation analyses, a vast amount of variation is accounted for by randomness, yet using a different approach yields a large effect size. Are all effect sizes going to be small, or is it the era of regression analysis that is causing this?

> Physical science models revolve around a small number of variables with large effects, so researchers can say this variable causes something else. But in biology, there are many outcomes, and any one of them can be influenced by a large number of variables that can change. It is simply a characteristic that researchers must confront. (Dr. Steve Grant)

Dr. Grant noted that that the physical and biological sciences are distinct in this respect. Physical science models revolve around a small number of variables with large effects, so researchers can say this variable causes something else. But in biology, there are many outcomes, and any one of them can be influenced by a large number of variables that can change. It is simply a characteristic that researchers must confront. Researchers in this field may need to develop a different approach to looking at the relationship between variables and outcomes.

Dr. Constable remarked that task-based fMRI has historically only looked at large effects. A preprint article by ABCD researchers features a figure showing rates of false positives versus false negatives as an effect of sample size. It is becoming more apparent that at a given false positive rate of .05, there is a very large false negative. For years, researchers have ignored many small effects because statistical power was lacking in these studies, but the field now has statistics that can examine these effects. Dr. Hancock asked: *As sample size expands, where do researchers decide to stop?* GWAS effect sizes can be taken to other layers (e.g., epigenetics) or the effects of genetics on neuroimaging is affected significantly by the developmental stage and other factors, as well as measurement at a single time point. Researchers need to take all these factors into account.

Dr. Grant wondered whether the convention of statistical significance—that is, setting the p-value at .05 (developed 100 years ago)—should perhaps be changed. Perhaps the field should establish criteria for effect sizes to differentiate between what is acceptable for understanding biological meaningfulness versus clinical application.

Should the field adopt the convention that an effect size of .1 using Cohen's D is too trivial to worry about for clinical translation, but is acceptable for further exploration and research into biological mechanisms?

Dr. Ragini Verma asked: *Why would researchers make such a distinction?* Exploratory research can be conducted to point to clinical and biological meaningfulness, and investigators can set p-values to whatever they want. However, those who want to argue that an effect is biologically meaningful must establish some type of standard (i.e., replicable, etc.). An attendee argued that effect size is not a relevant variable. For any disorder of interest to this group, which are shaped by the combination of thousands of genes and environment, the detectable effect sizes are always going to be small. Even when effect sizes are large, they may not be clinically meaningful. One of the largest effect sizes seen in this field is the impact of alcohol dehydrogenase on alcoholism, which is not a good predictor of alcohol dependence. When this variable is removed, it does not change PRS. This variable is an example of one that would not be suitable for clinical discrimination, but is incredibly important for studying biology.

Dr. Constable noted that, with larger sample sizes, researchers will grow more comfortable with smaller effects. An attendee asked: *How can we best differentiate small population effects that are derived from small effects across the majority of the population from small effects in the population that are derived from large effects in a small proportion of the population?* Dr. Grant suggested that this is where exploratory graphical representations of data would be useful for identifying a cluster of observations at one tail of distribution that seems disproportionate. Such a situation may indicate that there is a small sample in a population affected more by variables. Dr. Hancock remarked that, generally, researchers do not yet know the context for a particular subpopulation. However, in genetics, some large effect sizes are only important for people of a particular ancestry (e.g., *ALD2* for Asians). There are ancestry effects and pan-ancestry effects, which do not reflect a difference in biology, rather an ancestry effect. GWAS can identify those effects, but they are negated due to ancestral differences.

Participants discussed which effect size measures to report. There are many options, including Pearson's r (for correlation) and Cohen's D (for group comparisons). Dr. Grant focused on the issue of standards for Type I error and whether researchers should establish standards for effect sizes. Clinical trial researchers seeking approval from the U.S. Food and Drug Administration must state an effect size for clinical meaningfulness a priori. One attendee was strongly opposed to this notion, as it would negate psychiatric genetic knowledge to date. Dr. Grant asked: *Is there any gene effect that is too small?* An attendee commented that no gene effect is too small for examining biology. Sequencing can indicate a rare variant that affects one or two families a great deal, but is washed out in the overall population. With larger population samples, researchers will observe both convergence and have the power to conduct subgroup analyses. Dr. Hancock agreed that there is no threshold in psychiatric genetics because the situation will vary greatly by phenotype. Currently, a positive result is five to 10 genes that emerge from GWAS. Dr. Constable commented that a great deal of science is exploratory, without *a priori* effect sizes, which are unique to drug studies with well-characterized variables.

An attendee remarked that the field's focus on individual genes is an issue. Rather, researchers need to understand pathways and combinatorial effects and discuss "systems" modeling. Dr. Grant noted that methods of complex systems analysis are well established, but these are not used much in biology. Should researchers use these methods? Dr. Constable commented that the move toward a systems approach is happening as the amount of data increases and with the emergence of machine learning. Dr. Hancock agreed that the use of data is not yet optimal. Dr. Max Owens suggested that solving the univariate effect size problem will allow researchers to apply concepts and questions to machine learning and systems approaches. That needs to happen because those approaches are even more unmoored from traditional statistics than bivariate regressions.

Dr. Grant commented that the classic NIH R01 grant is weighted toward confirmatory rather than exploratory research. Perhaps that emphasis should shift based on the current discussion. Dr. Hancock added that R01 research is designed to be lower risk. There is no need to eliminate R01s, but NIH may want to focus on other, more exploratory funding mechanisms. An attendee remarked that exploratory research does not have to be high risk, as gathering larger samples is low risk. In his view, the focus on hypothesis in R01 grant review has not been helpful, as it encourages researchers to guess at candidate genes that end up not being robust. In Dr. Constable's experience, grant reviewers are not concerned about studies that are underpowered. There are many indications that the approach to R01s may need to be revised.

Participants discussed small effect sizes in electronic health records. This area of research has very large samples, and researchers see many small but statistically significant effects. Should these effects be considered clinically meaningful? There seems to be a tension between the view that any reproducible effect that meets statistical criteria could be clinically meaningful and another that focuses on biological meaningfulness (because there is not enough knowledge for clinical meaningfulness). Dr. Grant asked: *How large would an effect have to be for one to declare it clinically meaningful?* Several participants agreed that there is no single answer to this question. The answer requires defining the distinction between clinical and biological meaningfulness. Both have to be held to the same standard. A variable with a small effect size could have clinical impact, but researchers need to consider how to expand on that finding. The same approach applies to biological impact. Dr. Hancock remarked that clinical meaningfulness has a bearing on both prevention and intervention. Researchers have identified factors that will be relevant for one but not the other.

Effects may sit on the edge of a nonlinear inflection point so that a little movement in one variable causes disproportionate movement in another. When is a nonlinear analysis justified in evaluating a small linear effect?

Dr. Grant focused the discussion on Malcolm Gladwell's concept of "tipping point." Researchers have found many small effects in genetics and population brain imaging. However, it is not clear whether they are meaningful in understanding biological processes or inform clinical practice or population health. In a prior session, a participant raised the point that a variable with a small effect in an observational study might produce a large outcome when experimentally manipulated. Nonlinear effects presumably explain this. Such effects happen in pharmacology, for example. He asked for participants' views on nonlinear approaches.

Dr. Hawrylycz asked for examples of nonlinear effects to frame the discussion. Dr. Grant mentioned chaos theory and the idea that small changes in initial conditions affect outcomes. Any phenomena that observes power laws will yield nonlinear effects. He asked: *Now that we are confronted with many small effects, will it be necessary to move to nonlinear analyses?*

Dr. Luna remarked that it depends on what is being studied. For example, brain imaging from childhood to adulthood is a reverse curvilinear effect. Dr. Grant asked whether researchers studying a developmental effect need a nonlinear term in their analyses. Dr. Luna stated that studying the stabilizing processes involved in the maturation to adulthood would require considering nonlinear terms. Dr. Grant noted that GLMs can have a nonlinear term, and he posed the question: *What is nonlinear analysis?* In the GLM, the nonlinear term is rough. Other models have embedded nonlinearity in the interpretation to a greater extent. Even in GLM, especially in brain imaging, people do not necessarily use a nonlinear term. A conceptual question: *Does nonlinearity help researchers better understand the effects?*

Dr. Gina Lovasi noted that given the increasing consideration of splines in imaging, researchers have grown more comfortable with specific time periods to have a better approximation. But there are times in which a threshold is useful. She argued that the field needs better measures. Dr. Grant commented that nonlinearity would allow one to understand how a very small observational effect could produce a much larger outcome effect. Dr. Lovasi suggested that this approach would be most valuable with nonmonotonic variables (e.g., nonlinearity with age).

Dr. King commented that one could argue that a small range of exposures in observational studies causes smaller effects, whereas there are larger effect sizes in controlled experiments. He wondered whether it is a question of the representativeness of exposures. Dr. Grant agreed that researchers have to account for the range of exposures observed. Presumably, population studies attempt to capture the range. Dr. King advocated for speaking about effect sizes in plain language that policymakers can understand rather than reverting to jargon. He suggested that researchers identify and adopt better ways to communicate concepts (e.g., risk) in ways that are meaningful and understandable. This is relevant to discussions of longitudinal data in which small effects aggregate over time to result in a large outcome.

Dr. Grant raised another issue about nonlinearity: dynamical systems require a nonlinear approach because variables wax and wane. One can address this with a linear model that changes over time in a dynamic system. In engineering, a goal is to keep things within a linear range, but he wondered whether ABCD researchers should deal with nonlinearities by assuming that subgroups drive them. Do researchers need to model nonlinearity? Dr. Palmer pointed out that there are models that address multiple subgroups. Researchers cannot necessarily identify the subgroups, but they can identify clusters and mixtures of distributions. An attendee commented that there are behavioral phenomena that are nonlinear (e.g., log-linear), and these might fit an exponential or hyperbolic model. The data could also be made binary and move to a GLM.

Dr. Lovasi remarked that if ABCD researchers are looking to nonlinearity to solve the problem of small effect sizes, some cautions may apply. For example, if the shape of nonlinearity is such that it has a ceiling effect, researchers will not be able to manipulate beyond what is observed. So often, these systems revert to homeostasis rather than respond to our manipulations. Dr. Grant concentrated on analyses that provide a better way of understanding either the influence of one or a multitude of small effects. For example, in PRS, some variables might be weighed more heavily. Alternatively, there may be a critical point at which an aggregate score changes the slope of the response. Dr. Palmer commented that researchers who study psychopathology or substance use observe odd distributions with super-zero inflations with the tails of distributions. When researchers try to normalize the data or use log-link functions, they are likely to lose the ability to find small effects. Researchers need to examine the tails of distributions. Dr. King stressed the importance of using zero-inflated models. An attendee suggested that a population study examining substance use (such as ABCD and HBCD) must control for the amount and timing of substance use, not only its presence or absence.

Dr. Grant commented that in other domains, researchers may not have the ability to quantify exposure to the same degree or in a binary space. *How would one apply a zero-inflated model in whole-brain data analyses given that most common imaging packages rely on GLM?* Participants discussed imaging packages and their assumptions and capabilities. Dr. Fair noted that permutation testing cannot be done properly because there are so many nested variables, as in the ABCD study. Dr. Tom Nichols and colleagues have developed a sandwich estimator and bootstrapping technique similar to permutation to address this issue. This approach is built into the ABCD study and is being put into the ABCD Data Exploration and Analysis Portal. Dr. Palmer remarked that moving toward generalized estimate equations will allow for random effects like family or nesting, as well as longitudinal effects. An attendee pointed out that zero inflation is particularly problematic in clinical instruments in which people report few symptoms. Dr. Grant wondered whether zero inflation is as much of an issue in neuroimaging as it is in clinical instruments. Dr. Holly Moore remarked that time is an important factor in nonlinearity (how many time points and the time scale on which the process being studied actually changes). Perhaps such issues could be addressed in future workshops or tutorials.

Dr. Grant asked participants to identify examples of a case for which using a linear analysis (regression) only accounts for a small amount of variance, but a higher order model turns it into a large effect because it accounts for more of the variance. Dr. Luna commented that executive function or inhibitory control from age 8 to 25 goes from linear to inverse nonlinear. One can account for variance more because there is a great deal of growth early on, which then stabilizes. Dr. Grant noted that there are probably cross-sectional examples as well. *When does one decide to move to a higher-order analysis when a small effect size is observed in a linear analysis? Should researchers conduct such an analysis as an exploratory approach?* Sometimes there is no obvious structure in population studies, so researchers fit a line to the data. *But when should they decide that a curvilinear approach is appropriate?* Dr. Lovasi commented that it depends on the independent variables. When one wants to address confounds, the approach is to fit the data very well and not worry about p-values. However, for exposures of interest in the ABCD study, there is more cost to losing simplicity of interpretation. This comes back to transparency and *a priori* versus *post hoc* exploration.

Dr. King suggested that researchers preregister a series of model testing, prespecify how they will make decisions among models, and label it as exploratory. In this case, preregistration is critical and probably necessary. Dr. Palmer remarked that it depends on the research question. If the emphasis is on prediction rather than interpretability, then a machine learning algorithm can determine the best model. The nonlinear approach will increase effect size, but a difficulty arises with interpreting those effects. When interpretability is important, then researchers need to preregister a linear versus nonlinear approach.

Covariates/Collinearity

Panelist: *Dr. Vince Calhoun*
Concept presenter: *Dr. Ragini Verma*
Facilitators: *Dr. John Matochik, Dr. Heather Kimmel, Dr. Kim LeBlanc*

Determining which covariates to include in statistical models is complex and nuanced, especially in large data sets. Given the impact of covariate selection on replicability and reproducibility, these decisions must be made thoughtfully. Are there optimal strategies for selecting covariate controls? What factors must be considered?

Participants agreed that there is no consensus on this important but difficult topic. Dr. Constable asked the following questions: *Which covariates are examined? Why do researchers study those variables? Do researchers want their models to be maximally generalizable?* Models based on large sample sizes can perform well when researchers do not account for all factors. However, covariates driving group differences is a problem, but this is not appropriate for individual subject models. Dr. Martin Paulus remarked on the complexity of covariates, and agreed that those driving group differences bring the analysis into question. Researchers should consider the variables not usually thought about but which describe subjects (e.g., education, income, parental factors) prior to analysis. Investigators should be clear on what their models are testing and the research question that the study will try to answer. Research questions dictate the inclusion or exclusion of covariates. Therefore, there is no single best way to determine covariates.

> Research questions dictate the inclusion or exclusion of covariates. Therefore, there is no single best way to determine covariates. (Dr. Martin Paulus)

Dr. Matochik asked: *Is it a good idea to have a standard set of covariates?* Dr. Constable noted two considerations: (1) whether the subjects are from a single site, and (2) the experimental parameters. An article from the United Kingdom Biobank reported that data from about 10,000 subjects found 200 covariates with some detectable effect. This points to the important issue of generalizability. Researchers should be aware of covariates, but they may not have to consider all of them. Dr. Verma pointed out that combining data requires harmonizing the measurement of covariates. There is not a one-size-fits-all answer, but researchers should keep covariates in mind when they conduct analyses. Dr. Owens remarked that it is impossible to select a uniform set of covariates to use for every study. For example, there are confounds between socioeconomic status and race.

Dr. Matochik asked whether collapse-ability (i.e., conducting the analysis with and without covariates) should be a requirement for publication. Dr. Constable stressed that there is no rigorous method for researchers to choose covariates. Dr. Owens suggested that researchers report bivariates as well as covariates to avoid misleading readers. This is important because covariates can change the direction of effects. An attendee recommended including a variable as a covariate if it shows a significant association with both the independent and dependent variable. This is a simple heuristic that recognizes that such a variable could be a confounder. Dr. Constable agreed that this approach makes sense with a small group of covariates (five to six), but remarked that additional covariates would make the analysis much more complex. The same attendee added that one should only include variables that correlated with the variable of interest as covariates.

Dr. Matochik encouraged participants to discuss assumptions about the number of covariates used. *Is there an optimal number of covariates to include in a study that meets all assumptions and still answer the research questions?* Dr. Calhoun commented that this process should be driven by the research question. Rules about degrees of freedom apply, but a few degrees of freedom can be lost if one has a large sample size. He suggested that researchers flag variables of interest to examine whether they change the relationships observed. An attendee commented that in large data sets, many factors can be related to variables of interest. This situation forces researchers to control for many variables. One can build a model and add variables with significant correlations, but where should researchers draw the line? Dr. Paulus distinguished between covariables and confounders. Covariables are of interest and influence the dependent measure. Confounds influence both the independent and dependent variables and set up relationships that are false. Confounding is a process rather than a variable, and researchers must determine its influence. An article by Yixin Wang and David M. Blei proposes a machine learning strategy to minimize the effect of confounding based on observed variables. However, Dr. Paulus noted concerns about confounding based on unobserved variables. Some researchers have considered the magnitude of confounding's influence needed to eliminate observed effects. A further complication is that different fields vary in their definition of confounding.

Participants discussed the best way to report covariates, including providing an alternative analysis. Dr. Calhoun focused on consideration of the research question and key confounders that might be related to it. From these considerations, researchers can design a primary model to test and then run additional supplemental analyses. It is crucial to ensure that the effects observed are not changing and that covariates do not eliminate effects of interest. Regarding accounting for interactions between covariates, he noted the importance of distinguishing covariates of interest from nuisance variables. With covariates of interest, researchers should be concerned about possible relationships with the main effects. Nuisance variables are acceptable as long as they are not associated with the variable of interest.

Dr. Matochik wondered whether large sample sizes negate the impact of nuisance variables. Dr. Constable argued for simple models that apply despite confounds and covariables. An attendee suggested that researchers should include some nuisance variables in their models (e.g., studying crime requires consideration of population density). However, analysis that controls for nuisance variables may yield significant effects that are not necessarily meaningful. Dr. Verma favored leaving room for exploratory analysis, as researchers are not aware of all relevant variables at the beginning of a study. Dr. Paulus remarked that an important step in analysis is to examine potential covariates in a model followed by any hypothesis-driven analysis. One way to avoid bias when changing the model is to conduct an initial analysis with different data, then test the model with a larger data set.

Dr. Palmer commented on the difficulty parsing confounders and mediators. When additional variance is general, there are more likely to be covariates. Selection of covariates is complicated by unknown confounding. Analysis can provide researchers with an idea of the role of those variables and estimates of the lower and upper bounds of the relationship between brain and behavior. One way to estimate a relationship is to use covariates to create an interval of where a true effect may be. Participants agreed that conducting an analysis with and without covariates (i.e., collapse-ability) is a good idea. Identifying potential covariates is more challenging.

Regarding reporting analyses with and without covariates, Dr. Lovasi distinguished between reporting variables with and without adjustment, and variables related to temporal factors. An attendee commented that researchers should report the statistics associated with covariate analysis at a minimum. Participants discussed whether reporting an analysis with all variables considered would have meaning. Dr. Palmer noted that when covariates are unclear (mediator or nuisance), then researchers should conduct analyses with and without. Dr. Dunn agreed that analysis with and without covariates should be included for transparency and added that researchers should make them available for downstream analysis. Dr. Palmer added that this issue is relevant for meta-analyses and appropriate comparisons. Dr. Lovasi remarked that some common covariates may downplay the findings of original studies that provided a clear view of relationships. Dr. Heather Kimmel commented that researchers need space to report different variables and present a complete picture.

Participants discussed how to account for the removal of a variable that affects others as a confounder. Dr. King suggested that researchers avoid interpreting covariate effects, as covariates perform a different job than other variables. Dr. Lovasi saw the task as one of optimizing choices. Researchers have multiple options for relevant variables and must consider their granularity and whether data is missing. Dr. Palmer emphasized the importance of the research question. Some important variables (e.g., socioeconomic status) cannot be considered covariates, as they both confound and have important effects that cannot be teased apart from main effects. Dr. Luna argued that if the research question does not relate to socioeconomic status, then there is no reason to include it. Dr. Palmer commented that considering associations that are not causal pathways yields a false effect. Rather, researchers should consider the entire context of variables, which is possible in the ABCD study. One can look at all confounders' potential influences, but does not need to control for them. Dr. Fair remarked that this is another version of confirmation bias.

Dr. Calhoun suggested that for a given relationship, researchers can evaluate whether other variables might have caused the observed effects. Researchers can also evaluate the size of an effect they may not have measured (i.e., the impact of unknown confounds). Dr. Lovasi suggested that researchers conduct a quantitative bias analysis, and Dr. King recommended developing specification curves. Dr. Luna remarked that machine learning can be valuable, as it allows researchers to consider all variables. Dr. King stressed the importance of determining the robustness of effects and region of confidence for each effect. Dr. Calhoun focused on reproducibility and recommended conducting an exploratory analysis on a subset of data, adjusting the model accordingly, and analyzing the rest of the data in an attempt to replicate the findings.

Dr. King remarked that covariates have different effects for outcomes that are not normally distributed, and not all disciplines are aware of this. For example, the effect of a covariate can depend on whether an individual is at high or low risk for a disease, which changes the parameters of interpreting results. Dr. Kimberly LeBlanc asked how participants approach socioeconomic status and race in their data analyses. Dr. Palmer noted that racial groups show different cortical morphologies, but such variations can be due to a host of social factors. Race and socioeconomic status can be confounding factors with no causal link. Therefore, it is important to control for them together because one cannot tease apart whether they are confounding or mediating. Dr. Lovasi noted that race is not inherently biological, thus observed brain differences could be due to social experiences. Dr. Palmer added that genetics does influence the shape and size of the brain in ways that have no bearing on behavior or cognition.

> Movement [during brain imaging] is a crucial issue, but there is no single way to approach it. (Dr. Vince Calhoun)

Dr. Fair commented that a neglected topic is that researchers do not often control for systematic artifacts in data. Dr. Calhoun agreed that this is an important issue that also applies in longitudinal studies, for which researchers must be careful to consider variables that change across experiments that are unrelated to measures of interest. Dr. Fair highlighted the problem of subject motion in the scanner (it relates to other variables and is difficult to control for), which researchers are not addressing. Dr. Calhoun noted that removing spikes related to head movement is one approach, but it is complicated by correlations with other variables and effects on measures of interest. Movement is a crucial issue, but there is no single way to approach it.

An attendee asked participants whether they use propensity score methods. Dr. Lovasi noted that these methods are appropriate for identifiable dichotomous exposures, but not necessarily continuous exposures.

Participants discussed site effects. One aspect of site effects is the difference between scanners. In this case, researchers can make the scanner serial number a covariate. For non-imaging data, site can be a covariate. Site effects tend to be 10 times smaller than family effects. Dr. Thompson remarked that socioeconomic status affects the interpretation of associations.

Sensitivity analysis (reporting the findings with and without a covariate) is a good way to address covariates. The analysis can determine whether a covariate changes data patterns in systematic ways. Half or a subset of the data can be used for the first analysis to better understand covariates. Then, researchers can apply the lessons learned to a second analysis (model training, cross-validation, sensitivity analysis).

Regarding sample size, Dr. Thompson commented that most scientific questions do not require 10,000 subjects. At a sample size of 1,000 to 2,000, most effects are stable. Smaller data sets are appropriate for a matched set of subjects (e.g., twins). What approaches are appropriate for addressing covariates in absence of a well-specified directed acyclic graph (DAG)? Dr. Thompson remarked that one approach is to classify covariates into three batches: (1) essential covariates that are needed for interpretation, (2) possible confounders that may be mediators, and (3) more speculative effects. This is a simple strategy that requires domain knowledge.

Dr. Matochik highlighted the importance of considering gene-environment interactions when selecting covariates. For example, with *in utero* exposure to nicotine, researchers can select variables related to the genetic predisposition to use nicotine as covariates and control for them. However, those variables are also mediators of effects. Dr. Thompson added that controlling for a mediator will bias effects on the low side. Dr. Matochik noted that the ABCD study has sufficient sample size to analyze data separately for males and females or control for interactions. Participants suggested conducting a stratified analysis by race/ethnicity.

Dr. Matochik encouraged participants to discuss how much information should be included when reporting the analysis results. Dr. Thompson focused on the importance of transparency about all steps of the analysis. Researchers should describe the process used to obtain the results (in supplemental materials) and make the code available for replication. Although prioritizing the results reported according to p-values seems like a bias, it is difficult to know what else could guide writing articles.

Participants discussed the utility of controlling for family-level exposures and possible deviations from them (e.g., confounds shared by twins). Dr. Thompson noted that although it makes sense to do so, family-level exposure could be a source of variation rather than a confound. Co-twins can act as a control for unmeasured variables. Participants discussed the distinctions between covariates and confounders. Dr. Thompson argued that “covariate” is a vague term for a variable that is not of primary interest to the researcher. It could be a moderator (as seen in interactions), mediator, or part of a causal pathway. A confounder has a causal link to the independent variable and a mechanistic link to the dependent variable. An issue is that a covariate can be a collider variable that is downstream from the independent and dependent variables. These characteristics make it difficult to determine the covariates to include in models.

Dr. Matochik asked whether participants report results with and without covariates as a standard practice. Dr. Thompson remarked that his group does report models with and without a covariate (i.e., conduct a sensitivity analysis to examine how the covariate changes the results). Dr. Matochik asked: *Are there sample sizes so large that confounds and nuisance variables do not matter?* Dr. Thompson stated that confounds and nuisance variables have nothing to do with sample size. Rather, they depend on the controls used in the research. Many of the variables that ABCD researchers want to control for are proxies for multiple factors, and this is the rationale to conduct a sensitivity analysis. Dr. Thompson remarked that studies always have measurement issues (e.g., confounds), which are not solved by larger sample sizes. Large sample size enables more precise estimates. Researchers can address random variation (e.g., measurement error, sampling variability)—but not systematic variability—with a large sample size. Variables that are not measured cannot be added to a model.

Regarding genetic ancestry in GWAS, researchers attempt to control for population stratification due to gene-environment variations. When they identify a variable that is correlated with all SNPs, they can analyze effects with principal component analysis. A technique in economics is to match pre-exposure trajectories and use that information as a counterfactual for the exposed group. Participants viewed a generic set of confounding variables with caution, as these depend on the scientific question. In all analyses, researchers should consult with a domain expert when considering covariates. However, some covariates are based on the ABCD study design. Researchers should consider these variables (site, family, race, sex, parents’ marital status, and household income), but these factors should not constitute a default set of covariates. Many variables not considered by researchers differ by study site, yet they can affect the results. Regarding the effect of family, participants endorsed using generalized estimates and controlling for this variable as a clustering random effect rather than a fixed covariate effect.

Exploratory vs. Confirmatory Data Analysis Frameworks

Panelist: *Dr. Jennifer Pfeifer*
Concept Presenter: *Dr. Erin Dunn*
Facilitators: *Dr. Vani Pariyadath, Dr. Janani Prabhakar, Dr. Stacia Friedman-Hill*

Distinguishing between the value of exploratory analyses (such as effect size estimation) and confirmatory, hypothesis-driven analyses is especially important for emerging areas of study. What approaches best inform theoretical constructs and models that don't require **a priori** ***hypotheses?***

Dr. Thompson remarked that exploratory analyses are often described as confirmatory, which causes confusion. There is pressure for researchers to obtain publishable results, and confirmatory results are not necessarily non-exploratory. Researchers find it difficult to characterize the entire analysis in journal articles. Dr. Palmer added that transparency is a problem. Exploratory studies are good, but they need to be reported. Associations must be transparent. Studies should not have only one hypothesis and look for a single association. Studies can involve multiple hypotheses and multiple associations. Dr. King mentioned the problem of theory vagueness, with researchers losing sight of the reasons underlying a hypothesis. He suggested that researchers should consider theories in a more specific and precise way.

Dr. Lovasi noted that neighborhoods are often considered in studies, and she wondered about variables that associate effect sizes with them. Dr. Luna responded that the null is difficult to prove. Hypotheses tend to be based on the current literature, and studies that do not confirm them lead to new hypotheses to test. Dr. Fair commented that the pressure to publish and obtain funding guides study design and the publication of studies with insufficient sample sizes. It seems the situation is unlikely to change until the professional funding pressure changes.

Dr. Janani Prabhakar asked about incentives to encourage preregistration. Dr. Dunn focused on the important role of journal editors, who are in a position to encourage preregistration. Knowing that top journals would publish null findings would take pressure off researchers. Perhaps journal editors might commit 10 percent of journal space to null findings. For the journal she edits, Dr. Luna remarked that a null finding is equivalent to a non-null finding. A commitment from editors is not needed, and a younger generation of researchers can present null findings in poster presentations. Dr. King, another journal editor, confirmed Dr. Luna's point. However, he argued that more needs to be done to ensure the publication of null findings. Dr. Dunn noted that some specialist journals will publish null findings, but the more general, larger journals have not followed suit. Dr. Fair commented that the scientific community needs to accept negative results in the same way as positive ones. The review of null results must be considered in funding decisions.

Dr. Luna remarked on the limitations to registered reports. Many researchers feel that they are too restrictive and that they take away the fun of discovery. However, there is room for transparency and discovery even after a registered report. Dr. Pfeifer mentioned the misperception that preregistered reports require too much specificity about how the analysis will be conducted. Detailed analysis plans can be extremely long, and science is constantly changing. Analysis plans can change in light of dynamic science and best practices, even in studies with preregistered reports. Dr. King added that confirmatory and exploratory analyses can be included in a preregistered report. Registered reports can be flexible, and the field needs to communicate the possibilities to researchers. Dr. Prabhakar asked: *How can NIH help in this respect?* Dr. Fair responded that registered reports are moving in the right direction, but are a small part of the picture. Small sample sizes can seemingly confirm hypotheses but still be wrong. The issues are multiple and complex.

Dr. Palmer explained that a registered report yielding a null or significant finding does not necessarily mean that it is not actionable. Collaboration and team science help this process. Dr. Luna added that small collaborations are encouraging, but laboratory studies remain essential. Transparency and replicability are key. For developmental cognitive neuroscience in particular, small studies have utility, but researchers need to be transparent and specific about the value of the findings. Big data can help replicate findings from smaller studies, but small laboratory studies drive innovation. Dr. Palmer agreed and added that collaborations between two or three laboratories can yield the same innovation.

> For developmental cognitive neuroscience in particular, small studies have utility, but researchers need to be transparent and specific about the value of the findings. (Dr. Beatriz Luna)

Dr. Prabhakar encouraged participants to discuss future frameworks. Dr. Fair commented that as cross-sectional studies yield the smallest effects, smaller samples can focus on specific questions. An attendee mentioned adaptive design (e.g., adding new arms to a proposed experiment), which allows researchers to change direction, save resources, and find more useful results. Dr. Fair remarked that in population-based studies, the terms of small sample sizes are different. So, researchers looking for small effects need large samples because subject variability is too high.

Dr. Prabhakar asked: *Should journals evaluate sample size and power for negative results?* Dr. Luna noted that she and her colleagues grapple with this. Researchers must delineate/verify appropriate power, and the steps cannot be too big. The field is slowly realizing that power is needed. Dr. Lovasi wondered how researchers can enable journals and other groups to review and verify the appropriate power. Dr. Pfeifer remarked that resources (e.g., funding) can be a problem if researchers were not originally funded to ask a particular question. Publication processing fees are also an issue.

Dr. Prabhakar brought up retrospective collaborations. Participants discussed the possibility of a centralized data science resource. Such a resource could standardize the methodology for experimental design and outputting data. NIH is now looking at reproducibility and centralized data, and inviting more participation in data science. Dr. King commented that many researchers outsource data/statistical analysis. The key question is methodology. The ABCD study could publish methods and procedures that would apply in 80 percent of cases to eliminate the need to rely on outside statisticians and their expertise. Participants agreed that guidelines to help researchers think through methodology would be helpful.

Dr. Prabhakar remarked that team science and collaboration raise questions about credit, funding, job placement, and other concerns related to professional and career development progress. Dr. Luna commented on the need for transparency about authorship in collaborations. Dr. Fair suggested that standardization in how researchers conduct their work could be a priority and could help facilitate distribution of effort. Dr. Palmer remarked that, ideally, the emphasis should be on results and collaboration rather than authorship and individuals. Dr. Luna commented that first and senior authors deserve more credit, and such matters are important for promotion and hiring. Dr. Prabhakar noted that letters of recommendation could mitigate that problem. Dr. Palmer added that more journals require authors to attribute specific tasks from a study (e.g., X wrote the manuscript, Y did this aspect of analysis).

What approaches best ensure valid and reproducible associations from a more narrowly focused hypothesis-driven analysis?

Dr. Prabhakar remarked that exploratory approaches are useful to emerging areas of study. She encouraged participants to discuss barriers to these approaches. In Dr. Thompson's view, exploratory approaches are being widely used but are sometimes mischaracterized as confirmatory. Journals may not favorably view articles that take exploratory approaches. Dr. Uddin pointed out that statistical trends reported in journals have changed over the past 10 to 20 years. Moreover, the statistical training of students in graduate school is different from those techniques reported in journals, so there is a need for education.

Dr. Prabhakar asked: *How do we distinguish between exploratory and confirmatory approaches?* Dr. Thompson noted that hypothesis-driven studies are confirmatory, and examining data constitutes exploratory research. Dr. Hawrylycz commented that "exploratory" or "descriptive" studies simply examine data. Dr. Thompson remarked on the nuances that can exist in research. For example, concentrating on a small part of a large data set is probably confirmatory, but could be considered exploratory. There are also distinctions between the types of statistical methods used for exploratory and confirmatory analysis. Dr. Robert Grossman noted that the terms “data driven,” “hypothesis driven,” and “hypothesis free” are also used to describe this work. Data-driven exploratory analysis can generate hypotheses as part of an overall process. An attendee commented that as an epidemiologist, she sees this debate in COVID-19 research. Epidemiologists are publishing articles based on big data sets, and they are appropriately using exploratory approaches. Critics say that hypotheses were not appropriately tested. Historically (as seen in the AIDS epidemic, for example) researchers initially explored how the disease behaves, which is appropriate. Causal inference methods should not be criticized at this stage of research. Dr. Thompson suggested that a way to distinguish between exploratory and confirmatory approaches is to write a directed acyclic graph (DAG). If a researcher can write a DAG and control for confounding factors, that is ideal. However, it is not possible for many studies.

> Nonconstructive criticism is counterproductive. The field should encourage researchers to share data and replicate findings as part of the scientific process. (Dr. Amit Etkin)

Dr. Etkin remarked that nonconstructive criticism is counterproductive. The field should encourage researchers to share data and replicate findings as part of the scientific process. Participants discussed the unfounded criticism that “confounding” cannot be part of exploratory analysis. Dr. Thompson commented on the usefulness of the literature on causal inference despite the need to be skeptical of it. An attendee advocated for the importance of exploratory work and generating hypotheses, as science should not be too restricted. Of course, inference and logic also have a place. Additionally, small effect estimates indicate that researchers do not yet understand the larger picture (e.g., it is difficult to interpret a weight ratio of 1:1). Dr. Etkin noted that it is almost impossible to avoid separate small samples. External samples are necessary because there are so many confounders.

Dr. Grossman focused on the accumulation of data over a period of time. The relationship between exploratory and confirmatory evolves over time. As more data accumulate, researchers can update their questions and hypotheses. Researchers should report data as exploratory and later as confirmatory with additional data and a new causal framework. He added that exploratory and confirmatory approaches represent stages of a process rather than a rigid distinction. Dr. Thompson stressed that scientists’ unwillingness to contradict their prior publications is a barrier. Dr. Etkin noted that the overall accuracy rate of any given idea is low. Dr. Lahey added that research should be self-correcting.

Based on concerns expressed in the meeting chat, Dr. Prabhakar encouraged participants to discuss domain knowledge. Dr. Uddin emphasized the importance of cross-disciplinary work. For example, collaborations between clinicians with specific disease knowledge and computer science experts on current techniques are necessary for optimizing machine learning. An attendee commented that domain knowledge experts need to inform data scientists, as they understand nuances of the data. Dr. Grossman stressed that domain knowledge changes in light of data and dynamic clinical contexts. Dr. Hawrylycz argued that a better way to store information than PubMed should be developed. Science now requires a more knowledge-based environment that enables researchers to compare data with current understanding. Understanding a complex biological system requires an elaborate information framework, whereas the current organizational system is focused on isolated published articles.

Dr. Prabhakar asked: *How can the field reduce barriers to exploratory approaches and how can NIH be involved?* An attendee recommended that NIH write Requests for Applications that invite exploratory research. Dr. Etkin suggested that researchers set aside a data set of about 20 percent of the sample for a later confirmation study. Dr. Thompson noted that this was considered for the ABCD study. Participants agreed that splitting the data can have value. Dr. Lahey remarked that data scientists conducting an exploratory study can come to naïve conclusions without input from domain experts. Dr. Thompson added that the field must be realistic about the contributions of domain experts. An attendee summarized that although domain experts are not always right, trans- and cross-disciplinary collaboration maximizes the likelihood of useful results.

Pre-specification of analysis strategies via registered reports can reduce researcher degrees of freedom and enhance transparency and reproducibility of results. What are perceived obstacles to submitting registered reports and what can be done to encourage it?

Dr. Pfeifer and her colleagues preregister all their neuroimaging studies. Doing so constrains analytical flexibility. A criticism is that registered report submissions yield initially insufficient detail, and she wondered what guidance could be put in place to make this initial step less onerous. Dr. Owens commented that it is difficult to balance changing paths in a way that is scientifically valid while adhering to the path you outlined in the preregistered report. Presenting exploratory data is acceptable as long as researchers do not represent these results as confirmatory of their preregistered plan.

Dr. Prabhakar encouraged participants to discuss the incentives and challenges. Dr. Pfeifer emphasized the importance of distinguishing between preregistration and registered reports. Preregistration does not guarantee publication, but it does lend some credibility or weight to one's methodology. Preregistration does not mean that one writes the entire article beforehand. Dr. Prabhakar noted a comment in the chat suggesting that preregistering and registered reports seem like writing the article before examining the data. Dr. Pfeifer responded that preregistration helps the researcher outline the methodology and identify unknowns. Preregistering and registered reports also benefit the field by placing restraints on one's degrees of freedom as a plan—i.e., laying out a course of action. Preregistering and registered reports do not preclude later exploratory work. Dr. Owens added that preregistration can record the important process of yielding null findings from an established analytical plan, which offers value. Dr. Pfeifer explained that journals oversee registered reports, whereas preregistration is not embedded in any particular publication. Researchers can preregister at any point, including even partway through a study. Registered reports have a more formal process and are associated with a particular journal.

Dr. Prabhakar asked about researchers' possible concerns regarding preregistration. For example, do they see it as a contract or obligation that restricts pursuit of an unexpected observation? Dr. Paulus recommended that NIH Institutes offer training opportunities on preregistration for researchers at various career stages. The ABCD study generated a standard outline for preregistering, which is a model that could be adopted on a larger scale. An attendee noted that regarding the open science framework and shared data, researchers are more willing to share data when funders, institutions, or journals make it mandatory. She acknowledged the potential resistance to such mandates, but argued that preregistration would encourage open science and shared data while permitting exploratory approaches. An attendee voiced the concern that a preregistration mandate might result in treating all experiments the same (as in GWAS studies). Dr. Owens remarked that in the GWAS example, preregistration could be sufficiently vague to offer little constraint. However, it would prevent researchers from identifying a gene *post hoc* and saying, "we found what we hypothesized." Preregistration can work well for simple analyses, but not for large projects involving analysis in multiple stages, which is a dynamic process. Analysis at one stage informs analysis at the next stage. Preregistration does not work for this type of methodology. Researchers face a difficult choice. Dr. Pfeifer argued that increasing transparency means that one can pinpoint the period of time analysis led to a change in the direction of research. Preregistration can help researchers identify those moments when they deviated from the plan, enhancing transparency.

Dr. Prabhakar raised the common concern that exploratory models involve overfitting data. She asked whether a grant analysis plan can replace preregistration. Dr. Pfeifer commented that a grant analysis plan can serve as an outline of a preregistration. However, researchers are limited to 12 pages in a grant application. For a large study, the analysis plan will be too large. For instance, the ABCD study published a protocol paper. Preregistration is only useful, to some extent, in its level of specificity. An attendee noted the ethical and statistical issues involved. Researchers can address overfitting the data in multiple ways, and other data can validate the results. Although it is important to follow all of the experimental/analytical steps, it is unethical for researchers to claim to have hypothesized a finding in a purely exploratory study. In his view, mandatory preregistration will not stop such unethical behavior, which is relatively rare.

Results can be valuable even if the results do not meet an arbitrary p-value.

Participants focused on the behavior of peer reviewers. Results can be valuable even if the results do not meet an arbitrary p-value. Adhering to this convention can lead to “p-value hacking” and other questionable practices. Dr. Verma remarked that journals can encourage greater structure than scientists want. It is acceptable and necessary to identify and correct a mistake, and too much structure prevents corrections. Participants commented that it is good that some journals will publish articles regardless of null findings. Researchers face a disconnect between results and identifying journals that will publish reports of these findings. An attendee noted the bias toward positive findings and that regarding publication, the statistical bar is much higher for negative findings. Dr. Pfeifer argued that registered reports can address this issue because the journal agrees to publish regardless of whether the findings are positive or null. Preregistered and registered reports can find small effect sizes, which can act as a disincentive to researchers and journals, some have speculated. In her team’s experience, registered reports with null findings still have a high impact factor because the registration lends weight to the negative results.