Journal Club Super Star

Adriane M. dela Cruz[1], MD PhD, Marisa Toups[2], MD, Lindsey Pershern[1], MD

[1]University of Texas Southwestern Medical Center, Dallas, TX and [2]University of Texas Dell Medical School, Austin, TX

Corresponding Author
Adriane M. dela Cruz
University of Texas Southwestern Medical Center
6363 Forest Park Road
Dallas, TX 75390-9119
Phone: 214-648-3741 Fax: 214-648-0167
adriane.delacruz@utsouthwestern.edu

Version 2.0

Updated June 2020

All changes from previous version noted in bold type.

UTSouthwestern
Medical Center

Contents

Introduction

Journal club, a gathering of colleagues to discuss a medical literature article, has been a part of medical education since the time of Osler, and the role of journal club in undergraduate and graduate medical education has been studied for more than 30 years. Journal clubs in graduate medical education typically serve dual roles of teaching skills in critical appraisal of the literature and keeping residents and faculty up-to-date on key findings. A small, early study suggested that journal club is not an effective way for psychiatry residents to learn critical appraisal skills [1], at least over a 12 week period in which the journal club format consists of resident-selected articles and a single resident leading the discussion of each article. More recent work has highlighted the importance of several elements for a successful journal club: utilizing a format that encourages the active participation of multiple residents [2], meeting monthly [3], clearly stating the goals of the journal club [3, 4], articulation of reasons for article selection for discussion [4], and emphasizing the connection of the article to clinical practice [3, 5]. The ACGME and ABPN resident training requirements incorporate the goals of journal club in multiple milestone sub-competencies including PBLI1, PROF2, PC3, PC5, MK1, MK3 [6]. Many programs, however, struggle to implement the recommended practices for journal club due to a lack of resources, including faculty who do not feel they have the skills necessary to structure and facilitate journal club.

To address these needs, we have created a new journal club curriculum, Journal Club Super Star. This curriculum consists of a set of primary literature articles paired with two reading guides: (1) a preguide that contains questions specific to the article and (2) a postguide that summarizes the study with commentary of the study results and design. The guides can be used by residents to structure their reading of the literature and by faculty journal club facilitators. The guides can be used effectively by any faculty interested in facilitating journal club, including faculty with less familiarity with research.

Goals and Objectives

- Increase resident knowledge of evidence base for psychiatric practice
- Increase resident knowledge of research design and statistics
- Enhance resident skills in critical evaluation of the literature
- Enhance resident skills in applying evidence to clinical practice
- Improve feasibility of faculty participation in journal club via provision of a standardized curriculum for use in a variety of programs, independent of faculty resources

Curriculum Overview

This curriculum seeks to unite the two primary goals journal club: enhancing knowledge for evidence-based practice and building skills in the critical appraisal of the literature. This curriculum currently provides materials for more than 30 individual sessions; we continue to add new materials each academic year as well as update topic areas when new, important evidence is published. In our institution, journal club is held approximately once per month, with 9-10 total sessions per academic year. Each session lasts 50 minutes and is held during protected didactics time. Residents are expected to participate in all sessions and to read and engage with all journal club materials prior to the session.

There are three documents for each journal club session: the preguide, the article, and the postguide. All materials are provided to the residents and faculty a minimum of two weeks before the session. The preguide contains a list of questions specific to the article to help residents engage with the article and to highlight aspects of the research design. The preguide emphasizes areas in which the authors made critical decisions in either the study design or the presentation of the outcomes. Each preguide contains a “technical point” that poses a specific question about statistics and design.

The preguides can be used as guides for the faculty facilitators. For most journal clubs, the 50 minute time frame will not allow full discussion of every question in the preguide. Faculty facilitators guide the session by selecting a subset of the questions in the preguide as the starting place for discussion. Facilitators are encouraged to ensure that all residents are active participants in discussion, which is supported by limiting the size of each group of residents. They are discouraged from assigning one resident to lead the discussion, as this can discourage preparation and engagement of the other residents.

The postguide provides a brief summary of the article, highlighting both the strengths and the weaknesses of the design and analysis. The postguide presents with an explicit answer to the “technical point” but otherwise does not directly answer the questions posed in the preguide. The discussion in the postguide ensures that consistency among information taught across all groups.

PGY2-4

Residents (PGY2-4) are divided into groups of 8-10, with approximately equal representation of each class in the group. Resident members and faculty facilitators of the groups are held constant through the academic year. In our institution, each group is facilitated by two faculty members, typically one clinician and one researcher. The clinicians who facilitate journal club are typically core residency teaching faculty. The PGY2-4 curriculum is designed to be held in a three-year cycle, with all residents reading each article over the course of 3 years, with some residents encountering the article as a PGY2, others as a PGY3, and others as a PGY4. This structure allows residents at different points in training to share reflections on the article together and allows the more senior residents to model skills in reading the literature for junior residents. It also fosters camaraderie across class years. The PGY2-4 journal club articles cover a range of topics and research designs. A sample 3 year cycle can be found in the Appendix.

PGY1

The interns meet on the same schedule as the residents, but the intern journal club contains only members of the intern class. This group is also led by two faculty members who are also authors of the curriculum (AMD and LP). The intern journal club focuses on reading large, randomized, controlled trials of medical interventions for common psychiatric illnesses in order to meet the needs of these learners who desire knowledge directly related to their daily clinical work. For interns, each trial is paired with a brief review article describing an aspect of research design and statistics to enhance resident knowledge in these areas and strengthen their abilities in careful examination of the literature. The *JAMA* essay series "Guides to Statistics and Medicine" (https://jamanetwork.com/collections/44042/guide-to-statistics-and-medicine) has been an excellent resource for statistics and design review articles. Each of these articles is typically 2 pages and provides information at the level appropriate for interns. *The New England Journal of Medicine* series "The Changing Face of Clinical Trials" (https://www.nejm.org/clinical-trials-series) provides longer (typically 10 page) reviews of important issues in research design. In our experience, these articles are typically above the level of most interns, though they are an excellent resource for faculty. We have chosen the strategy of utilizing these review articles in favor a statistics textbook as we have found that interns are much more likely to read the brief review and that the reviews present much more usable information than what it found in textbooks. Pairing the review with an important clinical trial makes the information in the review much less abstract and gives the interns the opportunity to directly apply the knowledge from the review article in the analysis of the primary literature article. A sample intern curriculum can be found in the Appendix.

Summary of Critical Elements

- Faculty selection of articles
- Consistent faculty facilitators throughout the academic year
- Expectation that all residents will actively participate in discussion in each journal club session
- Pre-guides provided to residents with articles to orient their reading
- Pre-guides use by faculty to facilitate journal club discussion
- Post-guides provide consistency across groups

Faculty selection of articles allows for a more comprehensive approach to journal club and allows us to build our journal club as a consistent experience across the PGY1-4 years. Faculty selection ensures that high quality articles with important findings are read and discussed for journal club. This allows us to balance the representations of diagnoses, treatment populations, intervention types, and study designs across both single academic years and across the entire journal club curriculum. We also seek to balance reading of older, foundational articles with new articles that highlight recent advances. The articles for an academic year are all selected prior to the start of the year with careful consideration of how each selected article fits into the overall curriculum.

Many programs with research faculty have researchers (either basic or clinical) or other subject matter experts facilitate the journal club session related to their area of expertise, with different faculty leading different sessions. This approach decreases faculty burden (any given faculty member is likely committing to leading only a single journal club session) and increases resident exposure to faculty with deep knowledge in a subject area. We have moved away from this approach in favor of having a consistent group of journal club facilitators; many (though not all) have facilitated each year since institution of the Journal Club Super Star curriculum. From a logistical point of view, consistent facilitators are much easier to manage, as dates and times for the sessions are provided at the beginning of the academic year and facilitators commit to attending all sessions. Setting expectations —both between the course directors and the faculty facilitators and between the faculty facilitators and the residents—is much easier with consistent faculty facilitators and insures consistency across sessions. The rapport between the facilitators and residents is also enhanced. Finally, with consistent facilitators, the faculty easily recognize trends in resident behavior across sessions and identify residents who are simply not reading the article ahead of the session or who need additional support for skills in evaluating the literature.

Many programs utilize a journal club model in which a single resident (or a pair of residents) is designated as the leader or facilitator for a session. We found that the designated leader was well-prepared for the session, but most other residents were not. We observed that the majority of resident session leaders did not know which parts of the article to emphasize such that key issues in the design or analysis were never discussed. We have found that residents are more consistently prepared and engaged when journal club is a discussion among all residents and the faculty facilitators. The pre-guides help residents focus on important parts of the article, regardless of comfort with research and the primary literature, so they can feel prepared for discussion and meaningfully participate. This helps prevent the discussion from being dominated by any residents more confident about scientific articles.

Prior to initiating the pre-guides, we found that the majority of residents did not know what to focus on when reading journal articles. Many placed over-emphasis on the Background or Conclusions, and many often skipped the Methods section completely. Many residents understood the statements in the Conclusions as definite truth, rather than seeing these as part of an argument that the authors make. The pre-guides explicitly seek to uproot this approach by encouraging the residents to question assertions and assumptions made by the authors. Residents are encouraged to consider whether the study population and outcome measures are appropriate for the research question being posed. The pre-guides also emphasize a careful evaluation of the data presented. Overall, use of the pre-guides empowers residents to engage with the literature.

The use of the pre-guides is also a benefit to the faculty, particularly given our decision not to rely solely on researchers or subject matter experts as facilitators. By highlighting important questions about the article, the pre-guides assist facilitators in managing the discussion; the questions in the pre-guide can be posed verbatim to the residents. The majority are true discussion questions and do not have a single right or wrong answer, and critically, do not merely ask residents to regurgitate information in the text without processing it. Facilitators can choose which questions in the pre-guide to emphasize during the journal club session. Given the unpredictable nature of journal club sessions, we have found it worthwhile to prepare and distribute journal club post-guides, which provide summary and perspective on the article and an explanation of the “technical point” in the pre-guide. Faculty who feel less prepared to answer the “technical point” can also refer to the post-guide.

Evaluation Tools

We use biannual program evaluations to assess domains of learning specific to the Journal Club curriculum. The residents complete a comprehensive evaluation of the didactic program components. The subset of questions related to the journal club are displayed below:

PGY1 Journal Club Evaluation	Poor (1)	Fair (2)	Good (3)	Very Good (4)	Excellent (5)
Value of Intern Journal Club (overall)					
Value in developing basic statistical skills					
Value in developing skills to understand research evidence					

PGY2-PGY4 Journal Club Evaluation	Poor (1)	Fair (2)	Good (3)	Very Good (4)	Excellent (5)
******Name of Journal Club Leaders******					
Value of Journal Club (overall)					
Value in developing skills in reading literature					
Value in developing psychiatric knowledge					
Value in developing life-long learning interest/skills					

Residents generally have a favorable view of the journal club. Since its implementation (2013-2014 academic year), ratings have consistently been between 3 and 4, with the current academic year scores as reported in the table below:

	Mean satisfaction score 2018-2019
PGY1 - Value of Journal club overall	4.07
PGY1 - Value in developing basic statistical skills	3.60
PGY1 - Value in developing skills to understand research evidence	3.87
[redacted]	[redacted]
PGY2-PGY4 - Value of Journal Club (overall)	3.44
PGY2-PGY4 - Value in developing skills in reading literature	3.52
PGY2-PGY4 - Value in developing psychiatric knowledge	3.54
PGY2-PGY4 - Value in developing life-long learning interest/skills	3.37

In addition, the ratings for intern journal club in its inaugural year, were much improved from prior scores for the “Basic skills of evidence-based medicine” course (Mean scores ranging from 1.8-2.4 in the prior 3 academic years).
The Milestones appropriate for journal club are: PBLI1, PROF2, PC3, PC5, MK1, MK3.

Adaptability

The curriculum is highly adaptable. Residency programs may use these materials for individual sessions or may utilize the full curriculum. Programs may choose to create new materials modeled on the materials presented in this curriculum to meet the needs of their program. The curriculum may be used as designed in the didactic setting or may be used for more spontaneous article reviews in clinical settings.

Innovation

The Journal Club Super Star curriculum is innovative in the use of the preguide and postguide to structure the discussion in the journal club sessions. It also includes emphasis on technical aspects of research design as well as encourages thoughtful discussion rather than article ‘fact-finding.’ These guides facilitate faculty in any setting to expertly guide resident discussions.

References

1. Fu et al. (1999) Is a Journal Club Effective for Teaching Critical Appraisal Skills. *Academic Psychiatry* 23(4): 205-209.
2. Rodriguez and Hawley-Molloy (2017). Revamping Journal Club for the Millenial Learner. *Journal of Graduate Medical Education* 9(3): 377-378.
3. Deenadayalan et al (2008). How To Run an Effective Journal Club: A Systematic Review. *Journal of Evaluation in Clinical Practice* 14: 898-911.
4. McLeod et al (2010). Twelve Tips for Conducting a Medical Education Journal Club. *Medical Teacher* 32(5): 368-370.
5. Hartzell et al (2009). Resident Run Journal Club: A Model Based on the Adult Learning Theory. *Medical Teacher* 31(4):e156-e161.
6. ACGME and ABPN. The Psychiatry Milestone Project. July 2015

Appendix 1: List of Articles, Alphabetical by First Author

Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.

Arnold LE *et al*. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE* 2015; 10(2): e0116407.

Aubry, T *et al*. One-year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 2015; 66:463-469.

Aybek, S *et al*. Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 2014; 71(1):52-60.

Billioti de Gage, S *et al*. Benzodiazepine Use and Risk of Alzheimer's Disease: Case-Control Study. *BMJ* 2014; 349:g5205.

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Caspi, A *et al*. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 2003; 301:386-389.

Clementz, BA *et al*. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* 2016; 173:373-384.

Cummings, JR and Druss, BG. Racial/Ethnic Differences in Mental Health Services Use Among Adolescents with Major Depression. *J Am Acad Child Adolesc Psychiatry* 2011; 50(2): 160-170.

Diav-Citrin, O *et al*. Pregnancy Outcomes Following In Utero Exposure to Lithium: A Prospective, Comparative Observational Study. *Am J Psychiatry* 2014; 171:785-794.

Donovan, NJ et al. Association of higher cortical amyloid burden with loneliness in cognitively normal older adults. *JAMA Psychiatry* 2016; 73(12): 1230-1273.

Dunsmoor, JE *et al*. Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 2019; 39(17): 3264-3276.

Harrison, NA *et al*. Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity. *Biol Psychiatry* 2009; 66:407-414.

LaFrance, WC *et al*. Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 2014; 71(9):997-1005.

Lee, JD et al. Comparative effectivenss of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 2017; 391(10118): 309-318.

Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.

Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Liu, R-J *et al*. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biol Psychiatry* 2012; 71:996-1005.

McGinty, EE *et al*. Trends in News Media Coverage of Mental Illness in the United States: 1995-2014. *Health Affairs* 2016; 35(6): 1121-1129.

McNiel, DE and Binder, RL. Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *Am J Psychiatry* 2007; 164: 1395-1403.

Miller, IW *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Olson, KR *et al*. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics* 2016; 137(3): e20153223.

Popova, V *et al*. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *Am J Psychiatry* 2019; 176(6): 428-438.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.

Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.

Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *N Engl J Med* 2006; 355(15):1525-1538.

Skoglund, C *et al*. Attention Deficit-Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biol Psychiatry* 2015; 77:880-886.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.

Telch, MJ *et al*. Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *Am J Psychiatry* 2014; 171:1091-1098.

Trivedi, MH *et al*. Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 2006; 354: 1243-1252.

Uher, R *et al*. Genetic Predictors of Response to Antidepressants in the GENDEP Project. The Pharmacogenetics Journal 2009; 9:225-233.

The UK ECT Review Group. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 2003; 361:799-808.

Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.

Weiss, RD *et al*. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. *Arch Gen Psych* 2011; 68(12): 1238-1246.

Wierenga, CE *et al*. Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biol Psychiatry* 2015; 77:642-652.

Wunderink, L *et al*. Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 2013; 70(9):913-920.

Yehuda, R *et al*. Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 2014; 171(8):872-880.

Yovell, Y *et al*. Ultra-low-dose Buprenorphine as a Time-Limited Treatment for Severe Suicidal Ideation: A Randomized Controlled Trial. *Am J Psychiatry* 2016; 173: 491-498.

Appendix 2: Sample Curricula for PGY2-4 and PGY1 Journal Clubs

Sample Three Year PGY2-4 Journal Club Curriculum

Year 1

The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.

Aubry, T *et al*. One-year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 2015; 66:463-469.

LaFrance, WC *et al*. Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 2014; 71(9):997-1005.

Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.

Wierenga, CE *et al*. Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biol Psychiatry* 2015; 77:642-652.

Diav-Citrin, O *et al*. Pregnancy Outcomes Following In Utero Exposure to Lithium: A Prospective, Comparative Observational Study. *Am J Psychiatry* 2014; 171:785-794.

Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Telch, MJ *et al*. Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *Am J Psychiatry* 2014; 171:1091-1098.

Skoglund, C *et al*. Attention Deficit-Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biol Psychiatry* 2015; 77:880-886.

Year 2

Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.

Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *N Engl J Med* 2006; 355(15):1525-1538.

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Uher, R *et al*. Genetic Predictors of Response to Antidepressants in the GENDEP Project. The Pharmacogenetics Journal 2009; 9:225-233.

Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.

Arnold LE *et al*. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE* 2015; 10(2): e0116407.

Olson, KR *et al*. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics* 2016; 137(3): e20153223

Liu, R-J *et al*. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biol Psychiatry* 2012; 71:996-1005.

Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.

Year 3

Weiss, RD *et al*. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. *Arch Gen Psych* 2011; 68(12): 1238-1246.

Wunderink, L *et al*. Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 2013; 70(9):913-920.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

Harrison, NA *et al*. Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity. *Biol Psychiatry* 2009; 66:407-414.

McNiel, DE and Binder, RL. Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *Am J Psychiatry* 2007; 164: 1395-1403.

McGinty, EE *et al*. Trends in News Media Coverage of Mental Illness in the United States: 1995-2014. *Health Affairs* 2016; 35(6): 1121-1129.

Cummings, JR and Druss, BG. Racial/Ethnic Differences in Mental Health Services Use Among Adolescents with Major Depression. *J Am Acad Child Adolesc Psychiatry* 2011; 50(2): 160-170.

Caspi, A *et al*. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 2003; 301:386-389.

Yovell, Y *et al*. Ultra-low-dose Buprenorphine as a Time-Limited Treatment for Severe Suicidal Ideation: A Randomized Controlled Trial. *Am J Psychiatry* 2016; 173: 491-498.

Sample Intern Journal Club Curriculum

For the Intern Journal Club, each primary literature article is paired with a brief review article that discusses an aspect of study design and analysis. For the final three sessions of the year, the Interns joined an upper level journal club.

Primary Article: Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.
Review Article: Pocock, SJ and Stone, GW. The Primary Outcome is Positive—Is that Good Enough? *N Engl J Med* 2016; 375(10): 971-979.

Primary Article: Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.
Review Article: Sox, HC and Lewis, RJ. Pragmatic Trials: Practical Answers to “Real World” Questions. *JAMA* 2016; 316(11): 1205-1206.

Primary Article: Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.
Review Article: Kaji, AH and Lewis, RL. Noninferiority Trials: Is a New Treatment Almost as Effective as Another? *JAMA* 2015; 313 (23): 2371-2372.

Primary Article: Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer’s Disease. *N Engl J Med* 2006; 355(15):1525-1538.
Review Article: Tolles, J and Lewis, R. Time to Event Analysis. *JAMA* 2016;315 (10): 1046-1047.

Primary Article: Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.
Review Article: Detry, MA and Lewis, RJ. The Intention-to-Treat Principle: How to Assess the True Effect of Choosing a Medical Treatment. *JAMA* 2014; 312(1):85-86.

Primary Article: The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.
Review Article: None—combined with PGY2-4

Primary Article: Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.
Review Article: None—combined with PGY2-4
Primary Article: Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Review Article: None—combined with PGY2-4

Appendix 3: Articles by Diagnosis

Some articles are listed in multiple categories.

Addiction

Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Lee, JD et al. Comparative effectivenss of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 2017; 391(10118): 309-318.

Skoglund, C *et al*. Attention Deficit-Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biol Psychiatry* 2015; 77:880-886.

Weiss, RD *et al*. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. *Arch Gen Psych* 2011; 68(12): 1238-1246.

ADHD

Arnold LE *et al*. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE* 2015; 10(2): e0116407.

Skoglund, C *et al*. Attention Deficit-Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biol Psychiatry* 2015; 77:880-886.

Anxiety Disorders

Dunsmoor, JE *et al*. Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 2019; 39(17): 3264-3276.

Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.

Telch, MJ *et al*. Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *Am J Psychiatry* 2014; 171:1091-1098.

Yehuda, R *et al*. Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 2014; 171(8):872-880.

Bipolar Disorder

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Diav-Citrin, O *et al*. Pregnancy Outcomes Following In Utero Exposure to Lithium: A Prospective, Comparative Observational Study. *Am J Psychiatry* 2014; 171:785-794.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

Conversion/PNES

Aybek, S *et al*. Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 2014; 71(1):52-60.

LaFrance, WC *et al*. Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 2014; 71(9):997-1005.

Dementia

Billioti de Gage, S *et al*. Benzodiazepine Use and Risk of Alzheimer's Disease: Case-Control Study. *BMJ* 2014; 349:g5205.

Donovan, NJ et al. Association of higher cortical amyloid burden with loneliness in cognitively normal older adults. *JAMA Psychiatry* 2016; 73(12): 1230-1273.

Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *N Engl J Med* 2006; 355(15):1525-1538.

Depression

Caspi, A *et al*. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 2003; 301:386-389.

Cummings, JR and Druss, BG. Racial/Ethnic Differences in Mental Health Services Use Among Adolescents with Major Depression. *J Am Acad Child Adolesc Psychiatry* 2011; 50(2): 160-170.

Harrison, NA *et al*. Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity. *Biol Psychiatry* 2009; 66:407-414.

Popova, V *et al*. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *Am J Psychiatry* 2019; 176(6): 428-438.

The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.

Trivedi, MH *et al*. Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 2006; 354: 1243-1252

Uher, R *et al*. Genetic Predictors of Response to Antidepressants in the GENDEP Project. The Pharmacogenetics Journal 2009; 9:225-233.

The UK ECT Review Group. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 2003; 361:799-808.

Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.

Eating Disorders
Wierenga, CE *et al*. Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biol Psychiatry* 2015; 77:642-652.

Personality Disorders
Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randmoized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Psychotic Disorders
Clementz, BA *et al*. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* 2016; 173:373-384.

Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.

Wunderink, L *et al*. Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 2013; 70(9):913-920.

SMI

Aubry, T *et al*. One-year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 2015; 66:463-469.

McGinty, EE *et al*. Trends in News Media Coverage of Mental Illness in the United States: 1995-2014. *Health Affairs* 2016; 35(6): 1121-1129.

McNiel, DE and Binder, RL. Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *Am J Psychiatry* 2007; 164: 1395-1403

Suicide

Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randmoized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Miller, IW *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

Yovell, Y *et al*. Ultra-low-dose Buprenorphine as a Time-Limited Treatment for Severe Suicidal Ideation: A Randomized Controlled Trial. *Am J Psychiatry* 2016; 173: 491-498.

Appendix 4: Articles by Intervention Type

Pharmacological

Billioti de Gage, S *et al*. Benzodiazepine Use and Risk of Alzheimer's Disease: Case-Control Study. *BMJ* 2014; 349:g5205.

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Diav-Citrin, O *et al*. Pregnancy Outcomes Following In Utero Exposure to Lithium: A Prospective, Comparative Observational Study. *Am J Psychiatry* 2014; 171:785-794.

Lee, JD et al. Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 2017; 391(10118): 309-318.

Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.

Popova, V *et al*. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *Am J Psychiatry* 2019; 176(6): 428-438.

Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.

Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.

Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *N Engl J Med* 2006; 355(15):1525-1538.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

Telch, MJ *et al*. Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *Am J Psychiatry* 2014; 171:1091-1098.

Uher, R *et al*. Genetic Predictors of Response to Antidepressants in the GENDEP Project. The Pharmacogenetics Journal 2009; 9:225-233.

Wunderink, L *et al*. Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 2013; 70(9):913-920.

Yovell, Y *et al*. Ultra-low-dose Buprenorphine as a Time-Limited Treatment for Severe Suicidal Ideation: A Randomized Controlled Trial. *Am J Psychiatry* 2016; 173: 491-498.

Non-Pharmacological

Aubry, T *et al*. One-year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 2015; 66:463-469.

Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

Miller, IW *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

The UK ECT Review Group. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 2003; 361:799-808.

Combination

Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.

Arnold LE *et al*. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE* 2015; 10(2): e0116407.

LaFrance, WC *et al*. Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 2014; 71(9):997-1005.

The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.

Trivedi, MH *et al*. Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 2006; 354: 1243-1252

Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.

Weiss, RD *et al*. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. *Arch Gen Psych* 2011; 68(12): 1238-1246.

Appendix 5: Articles by Population Studied

Child/Adolescent

Cummings, JR and Druss, BG. Racial/Ethnic Differences in Mental Health Services Use Among Adolescents with Major Depression. *J Am Acad Child Adolesc Psychiatry* 2011; 50(2): 160-170.

Olson, KR *et al*. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics* 2016; 137(3): e20153223

The TADS Team. The Treatment of Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psych* 2007; 64(10):1132-1144.

Adult

Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.

Arnold LE *et al*. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE* 2015; 10(2): e0116407.

Aubry, T *et al*. One-year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 2015; 66:463-469.

Aybek, S *et al*. Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 2014; 71(1):52-60.

Brown, ES *et al*. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *Am J Psychiatry* 2015; 172(10):1014-1021.

Caspi, A *et al*. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 2003; 301:386-389.

Clementz, BA *et al*. Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *Am J Psychiatry* 2016; 173:373-384.

Dunsmoor, JE *et al*. Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 2019; 39(17): 3264-3276.

Harrison, NA *et al*. Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity. *Biol Psychiatry* 2009; 66:407-414.

LaFrance, WC *et al*. Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 2014; 71(9):997-1005.

Lee, JD et al. Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 2017; 391(10118): 309-318.

Lieberman, JA *et al*. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *N Engl J Med* 2005; 353(12):1209-1223.

Linehan, MM *et al*. Dialetical Behavior Therapy for Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 2015; 72(5):475-482.

McNiel, DE and Binder, RL. Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *Am J Psychiatry* 2007; 164: 1395-1403.

Miller, IW *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Popova,V *et al*. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *Am J Psychiatry* 2019; 176(6): 428-438.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Raskind, MA *et al*. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry* 2013; 170: 1003-1010.

Sachs, GS *et al*. Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *N Engl J Med* 2007; 356:1711-1722.

Skoglund, C *et al*. Attention Deficit-Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biol Psychiatry* 2015; 77:880-886.

Song, J *et al*. Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *Am J Psychiatry* 2017; 174:795-802.

Telch, MJ *et al*. Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *Am J Psychiatry* 2014; 171:1091-1098.

Trivedi, MH *et al*. Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 2006; 354: 1243-1252

Uher, R *et al*. Genetic Predictors of Response to Antidepressants in the GENDEP Project. The Pharmacogenetics Journal 2009; 9:225-233.

The UK ECT Review Group. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 2003; 361:799-808.

Warden, D *et al*. The STAR*D Project Results: A Comprehensive Review of Findings. Current *Psychiatry Reports* 2007; 9:449-459.

Weiss, RD *et al*. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. *Arch Gen Psych* 2011; 68(12): 1238-1246.

Wierenga, CE *et al*. Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biol Psychiatry* 2015; 77:642-652.

Wunderink, L *et al*. Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 2013; 70(9):913-920.

Yehuda, R *et al*. Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 2014; 171(8):872-880.

Yovell, Y *et al*. Ultra-low-dose Buprenorphine as a Time-Limited Treatment for Severe Suicidal Ideation: A Randomized Controlled Trial. *Am J Psychiatry* 2016; 173: 491-498.

Geriatric

Billioti de Gage, S *et al*. Benzodiazepine Use and Risk of Alzheimer's Disease: Case-Control Study. *BMJ* 2014; 349:g5205.

Donovan, NJ et al. Association of higher cortical amyloid burden with loneliness in cognitively normal older adults. *JAMA Psychiatry* 2016; 73(12): 1230-1273.

Schneider LS *et al*. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *N Engl J Med* 2006; 355(15):1525-1538.

Perinatal

Diav-Citrin, O *et al*. Pregnancy Outcomes Following In Utero Exposure to Lithium: A Prospective, Comparative Observational Study. *Am J Psychiatry* 2014; 171:785-794.

Appendix 6: Articles by Study Design

Some articles are listed in multiple categories.

Basic science

Liu R-J *et al*. 2012. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biological Psychiatry* 71:996-1005.

Case-control

Aybek S. et al. (2014) Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 71(1):52-60.

Billioti de Gage S. et al (2014). Benzodiazepine Use and Risk of Alzheimer's Disease: Case Control Study. BMJ 349:g5205.

Caspi A. *et al* (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301:386-389.

Diav-Citrin, O *et al* (2014). Pregnancy Outcome Following *In Utero* Exposure to Lithium: A Prospective, Comparative, Observational Study. *American Journal of Psychiatry* 171(7):785-794.

Olson KR *et al*. 2016. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics*. 137(3): e20153223.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Skoglund, C *et al*. 2015. Attention-Deficit/Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biological Psychiatry* 77:880-886.

Song J et al (2017). Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *American Journal of Psychiatry* 174(8): 795-802.

Genetics

Liu R-J *et al*. 2012. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biological Psychiatry* 71:996-1005.

Caspi A. *et al* (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301:386-389.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.
Uher R *et al*. 2009. Genetic Predictors of Response to Antidepressants in the GENDEP Project. *The Pharmacogenetics Journal* 9:225-233.

Imaging
Aybek S. et al. (2014) Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 71(1):52-60.

Donovan, NJ *et al*. 2016. Association of Higher Cortical Amyloid Burden with Loneliness in Cognitively Normal Adults. *JAMA Psychiatry* 73(12):1230-1237.

Dunsmoor, JE *et al*. Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 2019; 39(17): 3264-3276.

Harrison NA et al. (2009). Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity.

Wierenga, CE *et al*. (2015) Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biological Psychiatry* (77):642-652.

Metanalysis
Arnold L.E. *et al*. 2015. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE*. 10(2): e0116407.

The UK ECT Review Group. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 2003; 361:799-808.

Observational
Cummings, JR and Druss, BG (2011). Racial/Ethnic Differences in Mental Health Service Use Among Adolescent with Major Depression. *Journal of the American Academy of Child and Adolescent Psychiatry* 50: 160-170.

Diav-Citrin, O *et al* (2014). Pregnancy Outcome Following *In Utero* Exposure to Lithium: A Prospective, Comparative, Observational Study. *American Journal of Psychiatry* 171(7):785-794.

Donovan, NJ *et al*. 2016. Association of Higher Cortical Amyloid Burden with Loneliness in Cognitively Normal Adults. *JAMA Psychiatry* 73(12):1230-1237.

McGinty, EE et al (2016). Trends in News Media Coverage of Mental Illness in the United States: 1195-2014. *Health Affairs* 35(6):1121-1129.

McNiel, DE and Binder, RL (2007). Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *American Journal of Psychiatry* 164: 1395-1403.

Miller, IW *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Olson KR *et al*. 2016. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics*. 137(3): e20153223.

Wunderink, L et al (2013). Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 70(9):913-920.

Randomized Controlled Trials

Anton, RF *et al*. Combined pharmacotherapies and behavioral interventions for alcohol dependence. *JAMA* 2006; 295:2003-2017.
Aubry, Tim *et al* (2015). One-Year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 66(5):463-469.

Brown ES *et al*. 2015. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *American Journal of Psychiatry* 172(10):1014-1021.

LaFrance Jr, WC *et al*. (2014). Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 71(9): 997-1005

Lee JD *et al*. (2018) Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 391(10118): 309-318.

Lieberman JA *et al*. 2005. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine* 353(12):1209-1223.

Linehan, MM *et al*. (2015). Dialetical Behavior Therapy for High Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 72(5): 475-482**.**

Popova, V *et al*. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *Am J Psychiatry* 2019; 176(6): 428-438.

Raskind M.A. *et al*. 2013. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry*. 170: 1003-1010.

Sachs, GS *et al*. (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine* 356(17):1711-1722.

Schneider LS *et al*. 2006. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *New England Journal of Medicine* 355(15):1525-1538.

The TADS Team. (2007) The Treatment for Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psychiatry* 64(10): 1132-1144.

Trivedi, MH *et al*. Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 2006; 354: 1243-1252

Warden D. *et al*. (2007) The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.

Weiss RD et al (2011). Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. Arch Gen Psych 68(12): 1238-1246.

Yovell, Y *et al.* (2016). Ultra-low-dose buprenorphine as a time-limited treatment for severe suicidal ideation: a randomized controlled trial. *American Journal of Psychiatry* 173(5): 491-498.

Translational Neuroscience

Aybek S. et al. (2014) Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 71(1):52-60.

Caspi A. *et al* (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301:386-389.

Clementz BA *et al*. (2016) Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *American Journal of Psychiatry* 173:373-384.

Dunsmoor, JE *et al*. Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 2019; 39(17): 3264-3276.

Harrison NA et al. (2009). Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity.

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Telch, MJ *et al*. (2014). Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *American Journal of Psychiatry* 171:1091-1098.

Uher R *et al*. 2009. Genetic Predictors of Response to Antidepressants in the GENDEP Project. *The Pharmacogenetics Journal* 9:225-233.

Wierenga, CE *et al*. (2015) Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biological Psychiatry* (77):642-652.

Yehuda, R *et al*. (2014) Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 171(8):872-880.

Journal club pre and post guides

The documents in this Appendix are the pre- and post-journal club guides used by our program from the 2014-2015 academic year through the **2019-2020** academic year, arranged in alphabetical order by first author. Over that time, the curriculum has evolved; in particular, the post-guides have become more thorough.

For a select number of articles, there are two sets of pre- and post-journal club guides. This is due to the differences in the intern compared to the PGY2-4 journal club structure, as detailed elsewhere in the curriculum.

Pre-Guide

RF Anton *et al*. (2006) Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence. *JAMA* 295:2003-2017.

Reasons for choosing this article

- One goal of journal club is give all residents the opportunity to read the major, large, randomized controlled treatment trials in the psychiatric literature. The COMBINE study is one of these trials and provides information on major approaches to the treatment of alcohol use disorder.
- The study lets us think carefully about the pros and cons of making multiple comparisons.

Background

- What do the authors emphasize about the prevalence of alcohol use disorders? Why do they emphasize the role of primary care?
- At the time the study was conducted (early 2000s), what was known about the role of medication and therapy in the treatment of alcohol use disorders? Which medications were approved for this indication?
- What do you think was the hypothesis of study? What was the research team hoping to accomplish?

Methods

- Who were the study participants?
- How many treatment groups were there? What was the control group?
- As used in the study, define the following: medical management, combined behavioral intervention (CBI). How did medical management and CBI differ?
- How was "alcoholism" defined for the purpose of the study (i.e., what were the inclusion criteria)? Does this seem reasonable?
- What were the study end points? How was relapse defined for the purpose of the study? How was "good clinical outcome" defined? Of all of these endpoints, which do you think is the most clinically relevant?

A technical point from the Results:

The beginning of the results section describes the number of study participants who remained in the study ("research retention") and the mean number of pills taken by the participants ("medication adherence"). Why are these things important to understanding the study results? In what ways could low study retention or low medication adherence make a study result invalid?

Results

- Review Table 1. Did the different groups show any baseline differences? Why is this important?
- What was the most common adverse event in the study? What do you make of this?
- What was the effect of naltrexone treatment on drinking outcomes? Effect of acamprosate? Of CBI? (See Figure 2 and Figure 4.)
- What were the findings with medical management? Do these findings support the role of primary care in treating alcohol use disorder?
- How were the study participants doing at the one year follow up?

Discussion

- What do you take away from this study?
- What treatment would you advise for a patient with alcohol use disorder? Would you recommend a specific medication, therapy, and/or a combination of interventions?
- If a friend who is a primary care doctor called and asked for advice on what treatments they should offer to patients with alcohol use disorder, what would you tell them?
- Do you think the authors made a wise decision to compare all of these interventions in the same study?

Post-guide

RF Anton *et al*. (2006) Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence. *JAMA* 295:2003-2017.

Take Home Summary

This article is a randomized clinical trial comparing several different interventions for adult patients with alcohol use disorder. The central problem the investigators wanted to address is the treatment of alcohol use disorders in outpatient settings without comprehensive substance abuse services, which is the setting in which these patients typically present. This study, which was conducted about 15 years ago, remains important due to its large size and completeness of tested interventions. The number of comparisons conducted, however, can make the data difficult to interpret (in fact, this difficulty made some faculty members hesitant to include the article in journal club; other faculty felt this was a strong reason to have us read and discuss this paper). At the time the study was conducted acamprosate had not yet been approved by the FDA, and previous studies of Naltrexone looked only at people already receiving CBT for alcohol use disorder. The COMBINE authors were thus interested in an important effectiveness question — is true addiction specialty care with medications and rigorous psychotherapy necessary for people with uncomplicated alcoholism?

Thus, the goal of the study was to determine the efficacy of each medication alone and in combination, in the setting of the type of medication management typical of primary care compared to therapy (combined behavioral intervention (CBI)). The authors also wanted to test if there was a medication-specific placebo effect, so they included a group that received CBI and did not take any pills. *This led to a total of 9 groups*. To power the study adequately with so many comparisons, over 1300 subjects were enrolled across 11 sites. The study excluded participants who required medication for any other psychiatric diagnosis or who met criteria for another use disorder in addition to alcohol. Participants were expected to take 8 pills per day (pills were a combination of naltrexone, acamprosate, and their respective placebos). They were followed to a primary endpoint at 16 weeks with a follow-up endpoint of one year. The authors utilized 2 co-primary efficacy outcomes: (1) percent days abstinent and (2) time to the first heavy drinking day; they also include several secondary outcomes, including a composite “good clinical outcome” at the end of treatment and drinking outcomes at 1 year follow up.

The authors imply (but don’t explicitly state) that their primary hypothesis was that each treatment separately would be superior to placebo. Table 4 and Figure 2 present a composite of the study results at the end of the initial treatment period. First the authors examined main effects of the three treatment (Naltrexone, Acamprosate, CBI) compared to their controls – this means they compared all the subjects who got the treatment to all those who did not, without accounting for the other interventions – and found there was no main effect of any treatment on percent days subjects remained abstinent. There was a significant main effect of naltrexone on time to a heavy drinking day compared to placebo when including only those subjects who had at least one heavy drinking day during the 16-week study period. There were no main effects of acamprosate compared to placebo, or CBI compared to no CBI on heavy drinking days.

Interestingly, the most significant effect was the interaction between naltrexone and CBI on percent abstinent days. In when comparing CBI with no CBI, those with the worst outcome (75.1%) had neither (placebo + no CBI) while groups that got either or both were about the same (that is no benefit was seen from the combination, over one or the other. The group with the highest percent of abstinent days reached 80%, so the absolute difference was equivalent to a small effect size of about 0.2. A similar result was found for "good clinical outcome" (essentially, at least partial remission). At one year, no intervention met the significance threshold of 0.025, but the 16 week finding of naltrexone came in close at 0.04.

There are many, many ways to interpret these data. It may feel tempting to throw up your hands and run away from trying to make sense of this paper (the authors also probably regretted some decisions made in the study design which was not ideal to test for main effects). Be reassured that naltrexone and naltrexone+acamprosate prescribed in the primary care setting or in a general psychiatry setting without the time and resources for therapy are likely to be beneficial to patients. As you think through these data, keep in mind which interventions are being compared, at which time point, and which outcome is being measured. One major reason why the outcomes may be blurred despite the number of groups is that the study selection process may have interfered with good outcome by preventing subjects from taking adjunctive psychiatric medication during treatment, and by requiring a difficult regimen of many pills each day. The CBI arm appears to be called that because the requirements were too loose to call it therapy (there was no minimum number of sessions required) which may have weakened the effect of that intervention as well.

Regarding the technical point from the pre-journal club guide: Study retention and medication adherence are two critically important factors in interpreting study results. People drop out from studies because of medication side effects, burden from too many visits, or lack of benefit (i.e., drop out is not random). The same is true about medication adherence—typically, people have reasons for missing doses (e.g., side effects, lack of efficacy, pills are hard to swallow, it's too hard to remember three times a day). In other words, we can learn about whether the intervention was acceptable to the research subjects, and by extension whether it will be to patients. The other reason these numbers are important is the study power—if too many people drop out, the number remaining may leave the study under-powered. More difficult to detect is lost power from those who may stay in the study but aren't adherent to the medication. A nice rule of thumb is that you want to see 80% of the participants retained and 80% of mediation taken (or therapy sessions completed, etc). This study meets this metrics, which is reassuring because it lets you know that (1) taking naltrexone and/or acamprosate and attending therapy or medical visits are feasible for patients and (2) the retention and adherence were unlikely to effect the statistical analysis.

Pre-Guide

Arnold L.E. *et al*. 2015. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE*. 10(2): e0116407.

Reasons for choosing this article

- This article provides a meta-analysis of the overall benefits of treatment for ADHD. Many patients (and parents of patients) have questions about the benefits of such treatment, and this article may be useful for addressing these questions.
- This article uses the techniques of meta-analysis, and thus allows us to discuss the advantages, disadvantages, and potential pitfalls of this method.

Background

- The authors make clear that the goal of the paper is to examine the effect of treatment on outcomes and not symptoms. What is the difference? Why is this important? Related to this—in the methods, "ADHD symptoms" was not included as an outcome—what do you make of this decision?
- Why do the authors examine the effects of medication, non-pharmacological, and combination treatment? Why do you think they consider this a "primary interest?"
- What was the primary question of the analysis? What were the secondary questions?
- What would have been a reasonable hypothesis for this study?

Methods

- How did the authors identify published data to be included in the analysis? Why did they use multiple databases and search at different time points?
- What were the study inclusion criteria?
- What was included in the data extracted from each study? Was the appropriate information included? Is there anything that should have been extracted but wasn't?
- How was the effect of age at treatment initiation assessed?

A technical point from the Methods:

What is an effect size? How is effect size interpreted?

Results

- How many studies were included? Is this a big or small number?

- Describe Figure 2 in detail. (What's on each axis? What do the symbols mean? What's the difference between 2A and 2B?). In a sentence or two, summarize the findings that are presented in Figure 2.
- Which domains demonstrated the largest effect of treatment? Which treatment modality effected the most domains? Were there treatment modalities that effected some domains but not others? (See Figure 3)
- How did age of treatment initiation effect outcomes? Was anything about this result surprising?

Discussion

- What do you take away from this study?
- In clinical settings, what concerns about ADHD have patients/families raised with you? Using the data presented in this study, how would you address these concerns?
- The authors make an argument that their finding of no difference between short-term and long-term treatment is related to the timing of follow-up assessments. What is this argument? Do you agree with their reasoning?
- The authors state: "A limitation of this systematic review is the inclusion of studies of widely varied characteristics, for example, different study designs, study population types and numbers, types of informant or rater, follow-up intervals, diagnostic criteria, and treatments types." Is this a limitation or a strength? In what ways?
- What is publication bias? What steps do the authors take to address it? What statistical methods can be used to assess publication bias?

Post-Guide

Arnold L.E. *et al*. 2015. Effect of Treatment Modality on Long-Term Outcomes in Attention-Deficit/Hyperactivity Disorder: A Systematic Review. *PLoS ONE*. 10(2): e0116407.

Article Summary

This is a meta-analysis of treatment outcomes for Attention Deficit Hyperactivity Disorder that takes an interesting and very clinically relevant strategy. Given the potential serious impact of ADHD on children through the lifespan, as well as potential side-effects and risks of stimulants, many families understandable want to know the strengths and weaknesses of the two main treatment options, medications or behavioral therapies. Articles like this one have made a substantial contribution to the ability of clinicians to have that conversation with families by focusing on functional outcomes that both adult patients and parents of children care about.

The authors used a set of guidelines for meta-analysis called PRISMA that are designed to help readers have faith in the results since meta-analyses are highly subject to biased results if the criteria for study selection are biased intentionally or unintentionally. The primary purpose of the PRIMSA guidelines is to increase transparency of reporting for better evaluation. The methods relied upon here do not require access to the original data but instead use summary statistics to infer cumulative effects from many studies. They looked at two types of "success:" the fraction of studies that had any statistically significant result, and the effect size. The first looks at whether or not a treatment is likely to work and the second at how much improvement is likely.

One of the more interesting results of this study is that while behavioral treatments were more likely to work, they resulted in less improvement than medications. It shouldn't be surprising then, that the combination of the two was superior to either treatment alone. That said it was interesting how few (that is zero) studies had effect sizes versus placebo/control for combination treatment. Given the number of included studies overall however, these results are fairly robust.

When examining individual outcomes however, the number of studies drops substantially with a particular deficit for behavioral and combined treatments. The two outcomes with the best data are academic and social functioning and these also support the main finding that combination treatment is superior to either modality alone. This pattern is also implied for the other outcomes that had all three types of trials. Secondary analysis of age of treatment and age of initiation showed little effect. Finally they examined the length of follow up or length of total treatment; this was notable for suggesting (though not clearly proving) that benefit declines with time. They also found that adults had less benefit that children and adolescents from all treatment modalities.

Comments

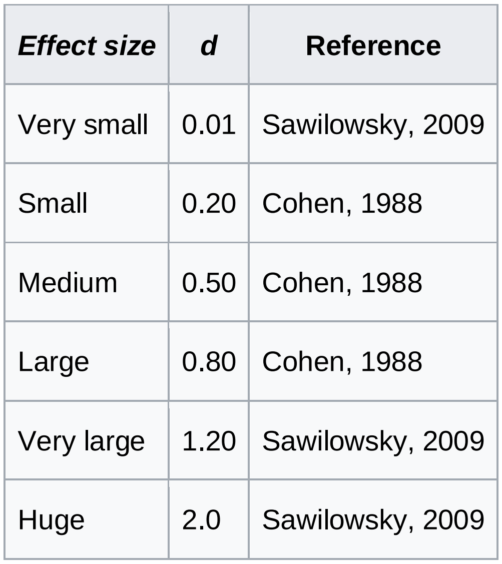
ADHD can be one of the more challenging disorders to assess and treat given the difficulty of establishing norms for neurocognitive function over the lifespan, given the large number of settings in which patients may have to function. This analysis supports that treatment for ADHD is generally more successful than treatment for other neuropsychiatric diseases in terms of functional outcomes –

comparable data for schizophrenia, for example shows very little benefit to function, especially from medication.

On question this meta-analysis leaves open is that of relative value of treatment as children mature into adults. Other reviews (referenced here) have suggested that ADHD treatment while very successful in the short term may not provide long term benefit. Is this because academic and later occupational challenges increase as patients grow into adulthood? Because medication tolerance develops? The authors also note that for some outcomes this pattern is inverted – particularly for impulsive/addictive behaviors, for which adults had much better outcomes. This would imply a fairly strong effect of context/environment on outcomes. The authors acknowledge that the inability to capture much of this information may effect the results. Meta-analyses that provide access to the full data sets of the original studies can go part of the way towards correcting this deficit.

Technical Point

Effect size is a way of standardizing the change in outcome variables so that we can more easily compare outcomes from studies. Additionally it helps us recognize when a statistically significant finding is “too small to care about” in practice. Increasingly experts in clinical research statistics recommend inclusion of effect sizes in addition to, or even instead of, significant testing. The table below (cribbed from Wikipedia) shows how to interpret effect size values for Cohen’s d, which derives from the comparison of two distributions (e.g. a t-test). Note, that much like the use of p <0.05 for significance, these values are an expert guess at the meaning of a mathematical result.



Effect size	***d***	**Reference**
Very small	0.01	Sawilowsky, 2009
Small	0.20	Cohen, 1988
Medium	0.50	Cohen, 1988
Large	0.80	Cohen, 1988
Very large	1.20	Sawilowsky, 2009
Huge	2.0	Sawilowsky, 2009

Cohen’s d is calculated by dividing the difference in means between the two groups by a pooled standard deviation. In meta-analysis like these, the standard error is often used because this takes the sample size into account as well, so that larger studies receive more weight than smaller ones.

This formula is also why the authors are able to make inferences about the effect across multiple studies with only minimal access to the information, and not the entire data set.

Cohen’s w, mentioned in this paper is a similar statistic that comes from a categorical analysis (e.g. remission vs no remission) such as a Chi-square test. Cohen’s f is used when the analysis has multiple groups (e.g. ANOVA).

Pre-Guide

Aubry, Tim *et al* (2015). One-Year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 66(5):463-469.

Reasons for choosing this article

- The article examines an important question in public psychiatry--is a particular type of housing service of greater benefit to patients with severe mental illness?
- There are many rotations in which you work with patients who are homeless and have severe mental illness, and this article lets us examine a non-pharmacologic intervention for these patients.
- The trial is randomized but not blinded, and it's important to consider in what ways the open treatment assignment may affect the results.

Background

- According the authors, what does "Housing First" describe?
- In your work with patients, what kinds of housing programs have you learned about?
- What is the authors' hypothesis regarding the efficacy of Housing First compared to a traditional housing program?

Methods

- Who were the study participants? How was "high need" defined? Jumping ahead to the data presented in Table 1, do you agree that the participants were "high need?"
- Why weren't participants/investigators blind to treatment assignment? Is this is a limitation?
- Which services were provided only to the intervention group?
- What were the outcome measures? Does the choice of outcomes seem reasonable to you? Is there anything else you would have liked the authors to measure?
- What do you make of the decision to provide ACT along with Housing First in the intervention group? How might this effect the outcome?

A technical points from the methods:

- On page 464, the authors state: "Site-level studies were powered to have a minimum of 65 individuals per group, allowing for the detection of a moderate effect size (effect size=.5) with a significance level of alpha=.05 and beta=.20, and anticipating a 25%-30% attrition rate." Why do the authors provide this information? How do you interpret/understand this sentence?

Results

- What do you make of the difference in attrition rate (the number of people initially recruited to compared to the number of study participants at each follow-up assessment) between groups?
- What are the major findings of the study? What effect did the Housing First intervention have on rates of homelessness? What effects did it have on the other outcomes?
- The authors report the effect sizes for several outcomes--how do you interpret these? Which of these effects are considered large?
- Did the type of housing intervention effect psychiatric symptoms or rates of substance use?

Discussion

- What do you take away from this study?
- Given that the study was conducted in Canada, do you think the results apply to patients living in Texas?
- Given that there was no effect of housing intervention on clinical symptoms, do you think that Housing First is an intervention that psychiatrists should care about for their patients? Why or why not?
- At the end of the article, the authors state "Our interim findings provide support for the redirection of programs and policies toward adopting Housing First to address chronic and episodic homelessness." Do you agree?

Post-Guide

Aubry, Tim *et al* (2015). One-Year Outcomes of a Randomized Controlled Trial of Housing First with ACT in Five Canadian Cities. *Psychiatric Services* 66(5):463-469.

Take Home Summary

This article describes the results of a large, open label, randomized trial of Housing First with Assertive Community Treatment (ACT) in adults with homelessness and serious mental illness (SMI). People who are homeless and have SMI often have other psychiatric and medical comorbidities, and they tend to present at expensive health care sites (e.g. the emergency room) at higher rates and burden these systems because they tend to not receive care (or housing) that stabilizes them. Significant moral and economic arguments surround these patients with stigma often playing a sadly disproportionate role. “Housing First” describes intervention in which patients are offered immediate access to housing *and* treatment for SMI, substance use disorders, and other psychosocial services, without making housing contingent on adherence to the other elements. Typically, services are provided at the location of the housing as much as possible. This is a significant alternate to traditional housing programs, in which patients move from living in a shelter to transitional housing to permanent housing, based on meeting treatment/sobriety goals, and are often forced to leave programs if they are unable to adhere to treatment.

This study is a randomized open label trial of a housing first program with ACT services. The goal is to increase time participants are stably housed and to improve their quality of life. They also examined psychiatric and substance abuse outcomes, but, of note, didn’t examine financial outcome in this paper. The trial was conducted in 5 Canadian cities and participants were homeless adults with severe mental illness. Participants were stratified into high-need and moderate-need based on severity of psychiatric illness, number of hospitalizations, number of incarcerations, and presence/absence of a substance use disorder; the current manuscript describes the results only of the high need participants. Participants were randomly assigned to the Housing First intervention in which they received vouchers for the cost of housing, assistance from housing coordinators, and ACT services; participants had to agree to meet the conditions of their lease and meet with study staff weekly. A total of 950 participants were categorized as high need and randomly assigned to Housing First (n=469) or treatment-as-usual (n=481). At one year follow-up, approximately more than twice as many participants in the Housing First group (~70%) were stably housed compared to treatment-as-usual. After demographic adjustment, the odds ratio for housing was above 6, which is a very large effect. Housing First participants also demonstrated greater improvements in overall quality of life and measures of personal safety, leisure activities, and social skills. Both groups showed equivalent improvements in clinical symptoms and drug use. The study demonstrates that a Housing First intervention can significantly decrease homelessness even in study participants with severe mental illness and ongoing substance use--treatment for these conditions is not a prerequisite for maintaining stable housing, given appropriate support.

Technical points: The text contains the power analysis, typically considered a required part of a clinical trial and most other experimental research. The term 'power' refers the probability of an experiment detecting a 'true effect' in the population using a sample of a given size. Because you can't test the entire population, there's a risk of randomly sampling a group that doesn't reflect the population. Imagine that you are sampling rocks from a sack – 70% of the rocks are granite and 30% are marble. It's possible if you just pulled a few rocks, you'd conclude that the rocks were all granite, but as you pulled more rocks the proportion in your sample would approach the correct value more and more closely. Power dichotomizes that curve of how likely it is you would state correctly that there are more granite than marble rocks in the sack, based on the number sampled. The parameter used to represent this probability is beta – which for most studies is set at 0.2 or the inverse of 80%. The power depends not only on the sample size, but also the effect size. In terms of our example, this means that you need to sample more rocks to accurately determine that granite out numbers marble if the ratio is 55/44 than if its 70/30. Usually we set the effect size (in terms of points on a scale, for example) and the power, and then calculate the necessary sample size – and then call it a power calculation, which is somewhat confusing.

You might be more familiar with alpha – the probability of concluding that there is a difference using a given sample, when in fact none exists in your population. Usually alpha is set at 0.05 or 5%, and is the threshold the "p value" is compared to. Because power, effect size and significance threshold are linked, always be aware when reading papers that a significant p value doesn't mean much if the effect size is too low!

Pre-Guide

S. Aybek et al. (2014) Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 71(1):52-60.

Reasons for choosing this article

- The article uses modern neuroscience techniques to investigate a Freudian theory of disease and thus brings together two aspects of psychiatry that are often thought to be very far apart.
- Many of us have seen patients with possible conversion disorder (especially on Consults), but this disorder gets less coverage in didactics than some other disorders.
- The article serves as a way to discuss the pros and cons of the case-control design and the strengths and weakness of fMRI.
- This article is a bit difficult to read, in part because the authors use terms (like "escape event") throughout the results and discussion that are defined only in the methods.

Background

- According to the article, what was Freud's theory of the pathophysiology of conversion disorder?
- What areas of the brain are suspected to be involved in conversion disorder?
- What is the hypothesis of the study?
- Several brain regions are discussed throughout the article. DLPFC=dorsolateral prefrontal cortex; rIFC=right inferior frontal cortex; SMA=supplementary motor area; TPJ=temporoparietal junction

Methods

- Who were the study participants?
- The study uses a case-control design—why do you think that is? What are some reasons to use this design? What are the weaknesses?
- What tool was used to assess the presence of stressful life events? How was it rated? Any pitfalls in this process?
- What do the authors mean by "escape potential?" What do they mean by "escape event" and "severe event"?
- What task were the participants completing while they were undergoing fMRI? Why do you think the authors had each participant complete a task related to events that participant experienced rather than a more general task?

A technical point from the results (but related to statistics):

- Throughout the results, the authors describe finding “main effects” and “interactions.” What do these terms mean? For example, in describing the results of the reaction time task on page 55, the authors state “there was no main effect of group,” “no group x condition interaction,” “but a significant main effect of condition was found.” How do you interpret that?

Results

- Looking at Table 1—how well do the cases and controls compare? Were they matched on appropriate variables? Should they have been matched on other variables as well?
- The authors compare two groups (conversion disorder vs controls) under three conditions (severe, escape, neutral). If you are trying to understand how people with conversion disorder process painful memories, which comparisons are you most interested in? (For example, you could compare conversion disorder in severe condition vs conversion disorder in escape condition. You could also compare conversion disorder in escape condition vs controls in escape condition.) Which comparisons did the authors make? What were they hoping to understand by making those comparisons?
- What were the results with regard to reaction times (RTs)? Were reaction times different between those with conversion disorder compared to controls? What do you make of these findings?
- Did the authors choose specific brain areas to study? Based on what? (Hint: the authors describe doing both a “whole-brain analysis” and an analysis with “a priori regions of interest”—what’s the difference?)
- Which brain areas showed differential activation between patients and controls? Did this depend on the condition? How well do those match your expectation for the brain areas that were likely to be involved?

Discussion

- What do you take away from this study? What does the study say about neural processing of memories of life events in those with conversion disorder?
- Do the authors provide modern neuroscience evidence for one of Freud’s theories?
- Can you make a causal link from this work? Do you think that the patients process the life events in this way because they have conversion disorder? Or do you think they have conversion disorder because this is how they process life events?
- Are there other Freudian theories that you would like to see tested in this way?

Post-Guide

S. Aybek et al. (2014) Neural Correlates of Recall of Life Events in Conversion Disorder. *JAMA Psychiatry* 71(1):52-60.

Take Home Summary

This article describes the results of a functional MRI study of patients with conversion disorder and matched controls with the goal of testing Freud's idea that conversion disorder results when memories of stressful life events are repressed and converted into physical symptoms. The authors recruited 12 patients with diagnosed conversion disorder and 13 matched controls. Stressful life events were identified and characterized using semi-structured interviewed called the Life Events and Difficulties Schedule (LEDS). Each stressful life event reported by a study participant was rated by the authors on 2 variables: event severity and escape potential. Event severity was defined as the threat posed by the event, and escape potential was defined as the degree to which a subsequent illness would decrease the stress of the event. Events with high escape potential are thought to be most associated with the development of conversion disorder. Participants were included in the study if they had at least one severe event with low escape potential ("severe event") and one severe event with a high escape potential ("escape event"). The fMRI task required patients to answer true/false questions regarding three events they had experienced: a neutral event, a severe event, and an escape event (each participant completed the task three times: once for each type of event). The dependent variables were the reaction time to answering the questions and the brain areas activated during the task. Patients also rated how upsetting the task was to complete. In completing the fMRI data analysis, the authors subtracted out the activity level observed during the neutral condition from the activity during the severe condition and during the escape condition. They then compared each group (patients vs controls) in each condition (escape vs severe). All study participants showed longer reaction times for answering questions about escape events than neutral events, and there was a trend for a longer reaction time for severe events compared to neutral events. Interestingly, participants judged the escape trial to be less emotionally upsetting than the severe trial. Patients with conversion disorder showed more activity in the right supplementary motor area (SMA) and right temporo-parietal junction (TPJ) during the escape task than the severe task; controls showed the opposite pattern (more activation during the severe task than escape task). Patients with conversion disorder demonstrated less activation in the left hippocampus during the escape task than the severe task, which was not found in the controls. The authors argue these findings suggest that patients with conversion disorder activate motor areas and suppress memory areas when presented with information about escape type stressful life events—the type of events associated with the development of conversion disorder. The authors also looked specifically at activation of the dorsolateral prefrontal cortex, an area associated with executive function and voluntary memory suppression. This area showed the highest level of activity when patients with conversion disorder were presented with the escape condition, which is also interpreted as evidence of memory suppression. There are two other major findings of the study: 1) decreased activity in the right inferior frontal cortex in the patients compared to the controls in all conditions, which is interpreted as impaired emotional inhibition and 2) increased connectivity between the amygdala and SMA in patients in all conditions, which is interpreted as abnormal limbic (regulation of memory and emotion) motor connection. Taken together, the brain activation pattern in those with conversion disorder in response to remembering an event with potential for gain from illness supports the notion that conversion disorder is associated with memory suppression and differences in motor system processing.

The study has several important limitations. The authors made subjective judgments regarding which events in the participants' lives were associated with "escape" (becoming ill after the event could diminish stress/consequences and is thus likely to be associated with conversion) and which events were "severe" but did not have potential for escape. Because this is a case-control study, causality cannot be determined. It is unclear if the people with conversion disorder already had this pattern of life event memory processing prior to the development of the conversion disorder or if this pattern developed with or after the conversion disorder. There is also the larger question of how to interpret "activation" on fMRI—these data suggest the brain areas that could be involved but do not give any information about the neurotransmitters (which could be excitatory or inhibitory), receptors, or proteins that are underlie the differences in activity.

Regarding the technical point from the pre-journal club guide: The terms "main effect" and "interaction" refer to the interpretation of the statistical analysis, in which differences between groups are separately examined for each variable (main effect) and then analyzed with all variables together (interaction). The analysis makes comparisons on two different types of variables: group (patients with conversion disorder vs controls) and condition (severe vs escape). The "main effect of group" compares the data only on the group variable—this analysis compares the average reaction time of all of the patients compared to all of the controls, regardless of the condition. The "main effect of condition" compares the data only on the condition variable—this analysis compares the average reaction time of all participants during the severe condition compared to all participants during the escape condition, regardless of diagnosis. The "interaction" looks at both variables together to answer the question does the condition affect reaction time differently in the patients with conversion disorder compared to controls. To answer the question posed in the pre-guide, "the main effect of condition" means that reaction times were longer in one condition than the other; "no main effect of group" means that the reaction times of patients with conversion disorder were the same as controls; and "no group x condition interaction" means that the pattern of reaction time being longer for escape than neutral was true for both the conversion disorder patients and the controls.

Pre-Guide

S Billioti de Gage et al (2014). Benzodiazepine Use and Risk of Alzheimer's Disease: Case Control Study. BMJ 349:g5205.

Reasons for choosing this article

- This study attempts to answer an important clinical question: does use of benzodiazepines increase the risk for Alzheimer's disease? Psychiatrists prescribe benzodiazepines for a variety of indications, and we should thus be well-informed about benefits and potential risks of these medications.
- Reading this article lets us think about the best way of trying to answer questions about long-term risks of medications. What type(s) of studies can be used to answer a question like "do benzodiazepines increase the risk of Alzheimer's disease?"

Background

- Prior to this article, what was known about cognitive effects of benzodiazepines?
- The authors state "benzodiazepines might not cause the disease but rather be prescribed to treat its prodrome." What do they mean by this? Does this idea present a problem for the study?
- What (if any) hypothesis do the authors have for the study?

Methods

- What methods do the authors use to attempt to answer the question? Are there other approaches they could have taken?
- How do the authors identify cases, controls, and exposures? Do these seem reasonable?
- What are the upsides and downsides to use these sorts of large databases?
- Pay careful attention to the time period of benzodiazepine exposure that is being examined. Is it coincident with Alzheimer's disease diagnosis? Sometime before or after?
- The authors describe a long list of confounders. Why do they include these things? Should they have included more things in the list?

A technical point from the methods:

- The authors describe doing a sensitivity analysis. What is a sensitivity analysis? What does a sensitivity analysis test for? What is the goal of doing this? How does it differ from the primary analysis?

Results

- Is the study population big enough to answer the question?
- What is the main result of analysis—do the authors find an association between benzodiazepine exposure and diagnosis of Alzheimer's disease? How big is the association?
- Is there evidence of a "dose effect" for the association? How do the authors assess for this?
- Does a concurrent diagnosis of depression, anxiety, or insomnia affect the association?
- How do the results of the sensitivity analysis compare to the primary analysis? Does this affect your interpretation of the primary analysis?

Discussion

- What do you take away from this study?
- How do we separate correlation and causation? Do you believe there is a *causal* link between benzodiazepine use and diagnosis of Alzheimer's disease, or do you think this is simply a correlation and that some other factor is the causative link?
- The authors assert that the results are generalizable. Do you agree?
- Do patients need to be aware of these results? How would you approach that discussion?

Post-Guide

S Billioti de Gage et al (2014). Benzodiazepine Use and Risk of Alzheimer's Disease: Case-Control Study. BMJ 349:g5205.

Take Home Summary

This article describes a case-control study comparing the records of 1796 elderly Canadians diagnosed with Alzheimer's disease (AD) from 2000-2009 to the records of 7184 matched controls to assess for an association between previous treatment with benzodiazepines and the diagnosis of Alzheimer's disease. Data were taken from an administrative claims database that included information on diagnosis and prescribed medications; the medical records of individual patients were not examined, and no direct clinical assessment of patients was conducted by the study team. The authors found that that patients with AD were more likely (odds ratio 1.51) to have been treated with benzodiazepines in the 5-10 years *prior* to diagnosis than controls. There was no effect of low benzodiazepine exposure (defined by the authors as 3 months of benzodiazepine treatment, referred to in the article as 1-90 prescribed daily doses), but rate of AD increased with increased exposure, with OR of 1.33 for those with 3-6 months of exposure of 1.85 for >6 months. The type of benzodiazepine may affect the risk, as there was a numerically higher risk in those who took long-acting medications (OR 1.72) than short-acting medications (OR 1.43) (but note that the confidence intervals for these overlap). The authors make several arguments in support of the idea that these results reflect a causal association and not merely a correlation. They looked at benzodiazepine intake at least 5 years prior to AD diagnosis, which they argue is a long enough period of lag time that the benzodiazepines prescribed in the study are not merely treating non-specific, prodromal symptoms of AD that later manifest as memory impairment. The authors argue that the consistency between the primary and sensitivity analysis (which looked at benzodiazepine treatment at least 6 years prior to diagnosis) supports causality. They also argue that the observed dose effect supports a causal interpretation. They observed a similar rate of anxiety and depression in both cases and controls, so the authors do not believe that one of these illnesses (that could be treated with benzodiazepines) is the true causative factor for AD. The authors state that the clinical implication of their work is that benzodiazepine treatment should be limited to ≤3 months, especially in the elderly, to reduce the risk of AD.

Regarding the technical point from the pre-journal club guide: A "sensitivity analysis" is a re-analysis of the primary outcome made with a different set of assumptions to test the robustness of the findings. In this study, the sensitivity analysis looks at a different period of benzodiazepine exposure (6-10 years prior to diagnosis in the sensitivity analysis, 5-10 years prior to diagnosis in the primary analysis) to see if altering this assumption about the time period of exposure affects the results. In this case, the sensitivity analysis gives a very similar result to the primary analysis, which is generally taken as evidence that the primary analysis is a robust finding. A sensitivity analysis can be done in many different types of studies (randomized controlled trials, meta-analysis) as a means of testing if the observed result is dependent on assumptions made during study design or data analysis. For an expanded but straight-forward introduction to sensitivity analysis, see: L. Thabane et al (2013). A Tutorial on Sensitivity Analyses in Clinical Trials: the What, Why, When, and How. *BMC Medical Research Methodology* 13:92.

Pre-Guide

Brown ES *et al*. 2015. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *American Journal of Psychiatry* 172(10):1014-1021.

Reasons for choosing this article

- This article reports on a placebo-controlled trial of a novel medication for the treatment of cocaine dependence in patients with bipolar disorder. Given the lack of FDA-approved medications for cocaine dependence and the high rates of comorbidity between bipolar disorder and substance use disorders, this is a potentially clinically exciting finding.

Background

- Why do the authors choose to study participants with both cocaine dependence and bipolar I disorder? How might these patients differ from those with cocaine dependence without bipolar I disorder?
- What is citicoline? What was previously known about clinical effects of citicoline? What might be its mechanism of action, particularly regarding effects on cocaine use?
- What was the authors' hypothesis? What was the primary outcome measure?

Methods

- What assessments did the authors use to measure cocaine use and mood symptoms? Were these appropriate?
- What do you think about the decision to include CBT for all participants? (The authors provide additional relevant comments on the CBT protocol in the discussion.)
- How was treatment for bipolar disorder handled in the study? What medications were participants permitted to take? How was “stable medication” defined? Do these decisions seem reasonable to you?
- What do make of the decision to reinforce (with vouchers) attending study visits/providing urine samples regardless of UDS result?

A technical point from the Methods:

The authors state “the study was not powered for secondary analyses.” What does this mean?

Results

- What is the major finding of the study?

- In discussing the baseline data presented in Table 1, the authors report that there was a difference between groups at baseline in the mean duration of cocaine use. Do you think this has an impact on the outcomes?
- Describe the results as presented in Figure 1. There is a difference between groups early in the study, but this difference disappears at the end. What changes underlie this difference over time? I.e., did the citicoline group improve initially and then the placebo group improved later or something else?
- Why do the authors report information on missing data? What do you make of the amount of missing data?
- How do you interpret Table 2?
- What results are presented in Figure 2? What is "study survival?"

Discussion

- What do you take away from this study?
- The authors state that their data "suggest that citicoline might be most effectively used in an acute treatment to reduce cocaine use in inpatient settings while other treatments are initiated." Do you agree?
- On page 1018, the authors speculate on explanations for the results. Do you find any of these arguments particularly persuasive?
- Should we start prescribing citicoline to patients with bipolar disorder and cocaine dependence? What about patients with only one of these diagnoses?

Post-Guide

Brown ES *et al*. 2015. A Randomized, Double-Blind, Placebo-Controlled Trial of Citicoline for Cocaine Dependence in Bipolar I Disorder. *American Journal of Psychiatry* 172(10):1014-1021.

Article Summary

This study aims to determine efficacy of citicoline in preventing (decreasing?) cocaine use in patients with co-morbid bipolar disorder and cocaine dependence. A randomized placebo controlled design providing treatment for 12 weeks was used; the primary outcome was cocaine use as measured by urine drug screen. Patients were maintained on treatment for mood symptoms and changes in mood were monitored throughout the trial, and all participants received CBT. The design is notable for a system for maintaining participation in the study, in which subjects were given increasing incentives as they attended more appointments. The study enrolled 122 subjects who were randomized 1:1 into placebo or citicoline titrated up to 2000mg/day. The statistical analysis used a mixed model with intent-to-treat methods used to account for missing data.

Overall about 50% of the subjects in each arm had at least one positive drug screen over 12 weeks. Subjects receiving citicoline used cocaine significantly less than those on placebo. However, the model also showed that the difference between groups was greatest early in the study, with the rate of cocaine positive screens becoming nearly matched by week 12. There were no differences in side effects or changes in mood in either group. Drop-out did not differ between the two arms.

Comments

This study is important because it exams a clinically difficulty situation (comorbid bipolar disorder and cocaine dependence) that has not been extensively studied. Multiple features of the study design are attempts to overcome hurdles for this type of study. In particular, drop-out of subjects is a serious problem – in all psychiatric trials but in this population especially. High drop out causes several issues, but the largest of these is low power resulting from the sample becoming too small. Often, in an attempt to overcome this, previous studies were designed to be larger than necessary. However, funding for such “over-sized” projects was wasteful, and results from such studies may be skewed due to non-random dropout. Instead, this study attempted to recruit and enroll exactly as many subjects as indicated by the power calculation, maintain those subjects using incentives, and account for drop out using statistical adjustment.

The study results suggest citicoline may be useful in preventing (decreasing?) cocaine use in patients with comorbid bipolar and cocaine dependence. However, the effect seemed to wane over the course of the trial. There are several possible reasons for that. First, the dose of citicoline was titrated up over the first half of the study; it is possible that lower doses have greater efficacy, though mechanisms explaining this possible effect remain uncertain. It is also possible that citicoline shows rapid tachyphylaxis (“poop out”) so that sustained administration is not effective. It’s also possible that there are non-pharmacological reasons why over the course of the trial the citicoline group regressed to the “baseline” rate of use found in the placebo group. Regardless, these results are promising and suggest that further investigation into the use of citicoline is worthwhile.

Technical Point

Statistical power is one of the essential concepts of clinical science. In some sense “power” is a misnomer, because the calculation is really more about **minimizing the chances of probabilistic error**.

There are two types of statistical error called types 1 and 2. **Type 1 error** produces false positives and **Type 2** false negatives, though these terms are not synonymous – there are other causes of false positives and negatives. These errors are based on the reality that when you ask a research question about bipolar disorder you can’t practically test it on all people with bipolar; instead you **sample** that **population**. Type 1 error reflects the probability that your sample will randomly appear to support your hypothesis even though you would find it untrue if you could study the whole population. Similarly, Type 2 error reflects the probability that your sample would falsely find there is no effect, when one could be found in the entire population of interest.

Power depends on three factors: the **size the sample** (more is better), the **size of the effect** you want to measure (larger is better) and the **threshold you set for significance** (more permissive is better). These factors go into the calculation of a variable called “β” which is the type 2 error – the chance of getting a false negative result. **Power is (1- β).**

Researchers have varying degrees of control over these three factors. You usually can’t set the size of the actual effect, though sometimes you can know (or guess) what it is based on prior literature, or you can at least set a minimum bar – if a treatment decreased patients scores on a mood scale by only one point, would you care? You can also choose a significance threshold to some extent, though this value is traditionally set at p=0.05. You do, however, have control over the sample size (at least in theory – that’s why drop out is important!) and this is why so much of the conversation about statistical power centers on the number of subjects in a trial. As sample size increases, the sample comes to resemble the population more closely thus minimizing the chance of missing a real effect. Most clinical trials are designed for 80% power, and therefore set a β error rate of 20%.

Pre-Guide

A Caspi *et al* (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301:386-389.

Reasons for choosing this article

- There was debate about including this article in journal club. The reasons in favor of picking the article were that the finding of the paper—that a certain polymorphism in the gene for the serotonin reuptake transporter is associated with the effect of life stress on the development of major depression—is a major finding that you will see mentioned in textbooks, other articles, grand rounds talks, etc., and we thought it was worth examining the paper that contains the original finding. The article also presents an attempt to deal with the question of gene and environment interactions, taking us from nature vs nurture to nature and nurture.
- The arguments against choosing this article: (1) it's older (>10 years), so we are ignoring (for the moment) developments in the field since it was published, (2) it doesn't seem relevant to clinical practice the way a clinical trial does, (3) and it's very, very dense. The journal *Science* uses its own particular format, in which much of the methods are tucked into the results, figure legends, online supplementary material, or the references, and there are no headings labeling the introduction/results/discussion. Do not be discouraged if this article is a difficult read.
- Some methods details and some additional figures are available in the supplementary online materials for the article. Reviewing these is *optional*.

Background (approximately the first 3 paragraphs, until the sentence "we tested this G X E hypothesis . . .)

- What do you already know about the involvement of the serotonin (5-HT) transporter and depression?
- Where in the gene is the polymorphism of interest located? Is this a coding or non-coding region? What do the terms "short" and "long" refer to?
- What is the functional consequence of having an s allele? Theoretically, do you expect depression to be associated with one of the alleles?
- What (if any) hypothesis do the authors have for the study?

Methods (information is scattered throughout the article, some in results, some in figure legends, some in supplementary information found online)

- Who were the study participants? When were they assessed for stressful life events and MDD? What do you think of the decision to look at stressful life events only during the selected period of time?

- When and how was major depression assessed? Does assessing MDD at the chosen time point seem reasonable to you?
- What do you make of the attempt to get collateral information?

A technical point from the methods:

- In basic science models, how are gene x environment interactions studied? What factors are easier to control in animals than in people? What sort of controls do you need to assess a gene x environment interaction in people? What steps did the authors take to put these controls in place? In what ways do the need for controls limit the ability to generalize the findings?

Results (pg 387-388, from "we tested this G X E hypothesis" until the paragraph starting "until this study's findings . . .)

- How many participants experienced "stressful life events"? Did the rate of stressful life events differ by genotype group?
- How many study participants were diagnosed with MDD? Did this differ by genotype (see the legend for Fig 1 and 3)?
- In which genotype is there an association between stressful life events and depression? How are the groups combined/compared (i.e., do they compare homozygous for s allele vs homozygous for l allele or make some other comparison)? Does this seem reasonable?
- Is there an interaction between genotype and stressful life events on suicide?
- How does the analysis of childhood maltreatment and genotype interaction compare with analysis using stressful life events?

Discussion (starts at the very bottom of pg 388)

- What do you take away from this study?
- What does it mean to say that there is a gene x environment interaction?
- Is enough consideration given to events occurring before age 21? Diagnosis of MDD prior to age 26?
- The authors speculate about using these results to identify "those needing prophylaxis against life's stressful events." What do you imagine that would be?
- Do you think there are clinical implications of the study?
  Do you see a role for genetic testing in the clinical practice of psychiatry, now or in the future?

Post-Guide

A Caspi *et al* (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science* 301:386-389.

Take Home Summary

This “famous” article observes an interaction between a polymorphism in the gene for the serotonin reuptake transporter (5-HTT) and stressful life events on the risk of development of major depression. The polymorphism is located in a region of the gene that helps determine how often the gene is transcribed, the promotor. The different versions of the gene are designated the “short” or “long” allele, which differ by the number of repeats of a certain DNA sequence. The short allele is associated with less transcription (and thus less protein) and has long been thought to cause a meaningful difference in brain activity, though at the time this was published there was little direct evidence in humans. It didn’t seem that 5HTTLPR directly associated with depression. Because we already knew that stress was associated with risk of depression but that there is a lot of variability in the response to stress across humans, the authors wanted to know if the 5HTTLPR genotype affects depression indirectly, by affecting resilience.

The authors examined the 5-HTTLPR genotype, occurrence of stressful life events between ages 21-26, and the diagnosis of major depression in the past year at age 26 among participants in a longitudinal study of people living in New Zealand. For feasibility, all participants with at least one of the more rare s allele (s/s or s/l) were compared to those homozygous for the l allele (l/l). Among the 847 study participants, 17% were diagnosed with MDD at age 26 (58% of those with MDD were female vs 42% male). The authors found that participants with at least one *s* allele were more likely to develop MDD with increasing number of stressful life events. For example, with 0 or 1 stressful life event, MDD was equally likely among all participants. With 3 or 4 stressful life events, MDD was more likely among those with an *s* allele than those without an *s* allele (Fig 1B, Fig 3). The same pattern was true when examining the number of depressive symptoms, probability of suicide attempt, and collateral report of depression. The same pattern was observed when, instead of looking at stressful life events from age 21-26, the authors looked at the effect of childhood maltreatment. In those participants without childhood maltreatment, the likelihood of MDD was the same among all participants, regardless of genotype. However, in those with severe childhood maltreatment, MDD was more likely among those participants with an *s* allele (Fig 2). The authors propose that the genotype thus mediates the effect of stress on the development of MDD.

One major criticism of the paper comes from the design, which looked at stressful life events from age 21-26 and diagnosis at age 26. On the one hand although this design avoids dependence on less accurate historical self-reporting. On the other, evidence supports that stress in early childhood has a greater effect on depression risk than stress later in life. Although they also gathered such retrospective data about early life stresses, many people who develop depression may have already done so by the age window assessed in the study, making it hard to be sure these results generalize to most depressed patients.

Technical point from the pre-guide: Genetic research in behavioral disorders has a number of hurdles compared to that for other medical disorders. This paper was published 15 years ago and still stands out as one of the more successful genetic studies despite the battering the results have taken over the years.

Much genetic research starts with animal models, usually mice. Mice are unusual in that they can tolerate in-breeding very well. In fact, the mice used in research come from one of a variety of strains which though they may differ somewhat from each other, have essentially no genetic variation within a strain. So lab mice are, for all practical purposes, clones of each other. Or as scientists put it, they have the same genetic background (FYI, this is not true of the rats used in animal studies). That's why so many genetic experiments use mice – when a gene is altered for an experiment, the researchers know that the difference is the only one present in the genome. Under these circumstances differences in behavior caused by a gene variant stand out, and it's also easy to manipulate the experiences that animals have for comparison. Mice may be separated from their mothers for intervals close to birth, reared in "impoverished" conditions, or subjected to "bullying" by bigger, meaner mice to mimic the stresses of human life.

When it comes to humans though neither of these factors can be controlled with any certainty. Humans (thankfully) are allowed to breed as we choose for the most part, and we have no unified genetic background. In fact after this study was published several publications in east Asian populations suggested that the 5-HTTLPR might have the opposite effect in that population than it does in the Caucasian sample tested here. Several other studies that measured stressors or the onset of mood symptoms in different ways failed to replicate, or contradicted this initial result.

Additionally there are technical reasons why consistency is difficult in human studies. Most techniques we use for genotyping have some rate of error and we may miss important gene variants because we don't know how or where to look in the genome. Which scales are used to measure traumatic events and whether events are graded or even counted can make a big difference. In addition to difference between ethnicities in genetic background, cultural differences may affect how trauma is discussed, or even admitted. These days most scientists working in human genetics have a healthy skepticism of genetic studies like this one, and an even stronger skepticism of genetic tests that claim to have sorted this out to the extent of clinical usefulness.

Pre-Guide

BA Clementz *et al*. (2016) Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *American Journal of Psychiatry* 173:373-384.

Reasons for choosing this article

- This manuscript describes the initial outcomes from the Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) project, which introduces a neurobiological reconceptualization of psychosis. It offers a completely different way to think about disease diagnosis and classification.

Background

- Do you think of schizophrenia and bipolar disorder with psychosis as separate diseases or as part of a continuum?
- Several of the references in the first paragraph describe findings in cancer research, and the authors (in effect) argue that that was has been true for cancer may also be true for psychosis. Is this a fair analogy?
- What do you think was the hypothesis of study? What was the research team hoping to accomplish?

Methods

- Who were the study participants?
- How big was the study? Why is the size of the study important, given the research goals?
- What tasks did research participants perform? Why were these tasks utilized?
- What did the authors do with the results of each assessment? How did they go from the raw data to what was used to define the biotypes? Don't worry about the details—try to understand this broadly.
- In your words, how did the authors determine the best number of clusters? How did they test that the clusters were distinct from each other?

A technical point from the Methods:

In the Procedures section near the beginning of the Methods, the authors report "there were no site effects that influenced group comparisons on any laboratory biomarker measure." What are site effects? Why is minimizing (or eliminating) site effects particularly important for the B-SNIP research project?

Results

- What are biotypes 1, 2, and 3? What defines the differences between them?
- What do the terms “cognitive control” and “sensorimotor reactivity” refer to? Why are the biotypes compared on these dimensions in Figure 1?
- Describe Figure 2. Do biotype 1, 2, and 3 have any relationship to DSM diagnosis? To symptom severity? To function?
- How do the data from the biological relatives compare to the patients with psychosis (see Figure 1)? Does the data from the relatives support the biotype classification?
- What brain areas differed in patients with psychosis? Does the neuroimaging data support the biotype classification?

Discussion

- What do you take away from this study?
- The first sentence of the discussion states: “the neurobiological heterogeneity across the psychosis spectrum illustrates the difficulty with attempting to derive etiological and neurobiological distinctiveness from clinical phenomenology alone.” What do they mean by this? Do you agree?
- Based on the data presented, there are at least 2 ways to classify patients with psychosis: (1) by DSM diagnosis or (2) by biotype. How would you determine which is the better/more accurate classification scheme?
- The authors argue that their work may explain previous discrepancies in the literature on psychosis. What is their argument? Do you agree?
- What are the limitations of this study? How important are these limitations?

Post-Guide

BA Clementz *et al*. (2016) Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers. *American Journal of Psychiatry* 173:373-384.

Take Home Summary

Psychiatric nosology (i.e. the classification of disease) has a long history of uncertainty and disagreement. Traditionally most nosological systems, including the DSM, were developed using expert knowledge. Kraepelin's system of diagnoses may be the most famous example, but you are probably familiar with diagnostic concepts promoted by other famous early psychiatrists, including now reviled concepts like "hysteria." In the mid-to late 20th century the lure of increased computational power prompted interest in empirically derived groupings based on mathematical methods that subset large sets of data. Two commonly used methods, Factor Analysis and Principle Component Analysis, use concepts that you may have learned (and promptly forgotten) if you took vector calculus in college.

The Bipolar-Schizophrenia Network on Intermediate Phenotypes (B-SNIP) is a large multi-site project with a goal of examining psychosis across psychiatric diagnoses. Novel features include the size of the cohort (difficult given the low prevalence of these disorders), the breadth of the phenotyping, and the inclusion of not just healthy controls but also first-degree relatives of cohort members. The consortium has looked at a number of data types from this dimensional perspective; in this paper they report primarily on the results of neurocognitive testing. Some of this testing is similar to what you think of as IQ testing but much of it involves more basic examination of brain activity in response to stimuli using EEG measurement. The goal of this analysis was to identify which of these many measures can successfully divide patients from healthy subjects and then look empirically for subgroups that differ *within* the sample of patients, and to some extent, their relatives.

To understand what they did, focus first on the idea that the initial data steps were designed to reduce the number of variables used in the analysis – essentially you replace a set of variables that are mathematically related to each other with a single variable called a principle component. The reasons they did this are elegantly summarized at the bottom of page 375. After the number of variables was reduced, they used a clustering algorithm to put the patients into groups, using a couple of metrics designed to evaluate the quality of the clusters (primarily this is to justify the number of clusters; the algorithms can keep dividing the sample into more clusters but of course eventually this just becomes a mess).

They chose a three-cluster solution based on two types of data, which they call 'cognitive control' and 'sensoriomotor reactivity' which basically mean "how hard it is to stay on task with conflicting demands on your brain", and "how hard it is to filter stimuli." The authors refer to each cluster as a biotype, and the analysis revealed some interesting trends. First, biotype 1 contains more schizophrenia, and biotype 3 contains more bipolar, with biotype 2 having a more even diagnostic distribution. Biotype 3 subjects had the most normal scores in both cognitive domains while scores in biotype 1 were severely decreased in both domains. Biotype 2 showed increased rather than decreased reactivity, and moderately impaired cognitive control. When examining other types of data not used to generate the biotypes, they formed a clear pattern in which severity decreased from 1-3. Some evidence for a biological underpinning of the biotypes was suggested by the fact that the first-degree relatives tended to be the same biotype as their effected relative, though more modestly.

One of the major limitations of this type of study is that we can only enter data into it that we think to measure, so that often the results merely organize the input data. In this case the testing used assessed subjects with tasks that were known to show differences in people diagnosed with schizophrenia, which was seen as the prototypical psychotic disorder, so it is not surprising to see clustering of the sample by how “stereotypically schizophrenic” the subjects were. It also is typical to see clusters by disease severity –e.g. biotype 3 is consistently less severely affected. The places where data show a different pattern are likely to be more useful in terms of adding to our understanding of our patients. Here, the finding that some subjects had low scores in both cognitive domains (biotype 1) while others had high scores in reactivity (biotype 2) is one such finding that may prove helpful in selecting treatments.

Technical Point

A limitation of neuroimaging, genetics and other “big data” science projects is that the sample size needed to accommodate such “high dimension” data is very large (i.e. there are more genes and more brain regions than there are subjects, a difference we’d like to minimize). These studies therefore typically desire to include multiple sites so that as many subjects as possible can be recruited. Different sites may also have access to different ethnic populations, making a sample more representative. However, coordinating the activity of multiple sites is a huge challenge.

One confound in conducting research across sites is that each site may do things differently, leading to differences in the data. The ultimate concern is that the differences between groups are not due to true differences but are actually related to differences between how the sites conducted the study and collected the data. Sites may simply have staff with different training and habits, or may have historically backed rival scientists with meaningfully different views on how research in an area should be conducted. Sites may agree to use the same instrument and only find out after the study is underway that they were using different versions. Different brands of equipment also cause variation in data, as does using different labs. Versions of this problem cross scientific domains (e.g. the company you order mice from for cancer research can affect results), so the best researchers tackle it head on. Strategies for minimizing site effects include procuring the same (often expensive) equipment at each site and training staff from all of the sites together (usually at least once in person at more expense) to harmonize their methods and be sure they all knew which version of each assessment to use. Given the failure of many important research findings to replicate, this mundane work is more important than ever and is the mark of quality science.

Pre-Guide

JR Cummings and BG Druss (2011). Racial/Ethnic Differences in Mental Health Service Use Among Adolescent with Major Depression. *Journal of the American Academy of Child and Adolescent Psychiatry* 50: 160-170.

Reasons for choosing this article

- This article introduces the concept of health care disparities, with a specific focus of racial and ethnic disparities in care.
- This article allows us to think specifically about the barriers adolescents face to receiving care.

Background

- Why do the authors think it’s important to understand the rates at which adolescents receive care for depression?
- Why is it important to look at rates of care across different racial/ethnic groups?
- What gaps in the literature do the authors feel this study addresses?
- What hypotheses do the authors have for the study?

Methods

- Where did the study data come from? How was it collected?
- Related to the questions above: What is the National Survey on Drug Use and Health? Why is it conducted?
- How was it determined which survey participants had depression? What was the overall prevalence of depression? Did the rate of depression differ across racial/ethnic groups?
- How was treatment defined? What were the different types of treatment that the study assessed?
- How did the authors measure other things that may play a role in access to care, like family income and insurance status? How did they control for these variables in the statistical analysis?

A technical point from the methods/results:

- In the section of the methods describing the statistical analysis, the authors describe their procedures for “pooled weighted probit regressions and negative binomial regression models.” In the tables 2, 3, 4, several of the columns have headings referring to a “model” number. In your own words, what does the word “model” refer to?

Results

- What percentage of National Survey of Drug Use and Health participants identified as having a major depressive episode in the last year had received any treatment? What do you think of this number?
- How did race/ethnicity affect rates of treatment for depression? Which ethnic/racial group was least likely to receive care for depression? What do you think about these results?
- In table 1, what factors were NOT affected by race/ethnicity? What does this tell you?
- As you read the text of the results, carefully walk through tables 2, 3, and 4. What does RD refer to? What do positive numbers mean? What do negative numbers mean? What are the different models? Which factors increase likelihood of receiving treatment for depression? What factors make it less likely?

Discussion

- What do you take away from this study?
- How much of a role does family income and insurance status play in the lower rates of treatment among minorities?
- In describing the study limitations, the authors note "several constructs such as family status and health status are assessed with proxy measurements that may be imprecise." What do they mean by this?
- As a practicing psychiatrist, what should you do with information of overall low rates of care, with particularly low rates of care among ethnic/racial minorities?

Post-Guide

JR Cummings and BG Druss (2011). Racial/Ethnic Differences in Mental Health Service Use Among Adolescent with Major Depression. *Journal of the American Academy of Child and Adolescent Psychiatry* 50: 160-170.

This article addresses the question of how treatment for depression in teenagers varies by race and ethnicity in the US. This is an important question because in general our understanding of health and health systems is less well developed for children than adults. There is also specific concern that prevalence of (identified) mental health problems in youth is increasing and that our system lacks capacity to provide care to everyone, with the potential for systematic issues that contribute to racial and ethnic disparities in receiving care. This may represent a serious crisis for the next generation if care disparities aren't addressed.

The authors used publicly available data from the National Survey of Drug Use and Health, a large ongoing project in which Americans are administered detailed phone surveys to establish epidemiology of various disorders in the US population. Subjects are given a complete diagnostic interview and provide information about their family, economic and social situation, including insurance status. They are also interviewed about health care usage. The authors use survey data that described the care type (medication vs behavioral therapies), the care setting (inpatient, outpatient, or in school) and the type of provider. Data was used for subjects age 12-17 and care utilization was assessed for the year prior to the interview. Subjects were divided into five race categories: white non-Hispanic, Hispanic, black, Asian, and other, which was a mix of mixed race background and Native Americans and other small groups. It can be noted that these categories may not be very precise – putting Han Chinese and Hindu South Asians (not to mention Muslim South Asians) together in the same category may not be meaningful.

Using regression, the investigators first performed a weighted but unadjusted analysis to determine 1) the overall pattern of access to care by race and 2) which variables other than race seemed to have an effect on health care utilization. Table 1 shows the results of the analysis that included all the variables of interest across the 5 race categories. You can see there were significant differences in almost all the measures, with white race predicting more care. However if you look at the middle of the table where the overall assessment of need for services is found, you will notice that white were also assessed as having higher need than blacks or Asians (but not Hispanics). By eye, the differences in need seem smaller than the differences in access, and you can also see that there were differences in income, family structure, and insurance that may complicate the results. This is why the investigators performed the analysis shown in Figure 1, which is adjusted for those differences. Here you can see that whites receive the most care, followed by blacks and Hispanics, then Asians. The data also show that white teens are more likely to see a mental health specific provider and much more likely to take medication.

The results suggest that while Asians are the most economically similar to white and have higher rates of need, they are less likely to access care, while blacks and Hispanics seem to have care access that matches their socioeconomic status in pattern but cannot be fully explained by those variables alone. This points to different interpretations. Lack of care access in blacks and Hispanics may be driven by systemic racism more than in Asians, who may be driven more by factors such as cultural stigma that prevent parents from taking their teen children in to receive care. Of course without data capturing those variables, it's difficult to do more than speculate. Perhaps more importantly, the overall rate of service access was less than half! That means in general adolescent mental health needs in this country often go unaddressed.

Technical Point

We often use the term "model" as a verb to mean that a question of relationships between things has been answered using statistics. You may notice that this meaning is common in media and even in the scientific literature, but the term "model" has a more specific meaning. When performing an analysis that includes making a prediction about the overall relationship between constructs – not just the data we observe in a particular project, we use models to more narrowly define relationships. The simplest model is a straight line (i.e., a "linear model") and the majority of studies you look at will make the assumption that the relationships it examines fall in a straight line. Recall from high school math that this formula is y = mx + b, meaning that we assume that (for example) as age increases the rates of dementia increase at a steady rate m, with an adjustment factor (constant) of b.

Of course few relationships truly follow a line (for example, a line may work for ages 50-80; it probably won't for the whole human lifespan). So it's important to remember that 1) models are approximations, and 2) picking the right one matters. In this article things are especially tricky because the independent variable (race) has no clear form on its own (that is, for those who want a little stats jargon, it is a factor but is not ordinal) so the authors chose a fairly strange model. But that won't stop us from looking at a simple example that shows why choosing models correctly is important.

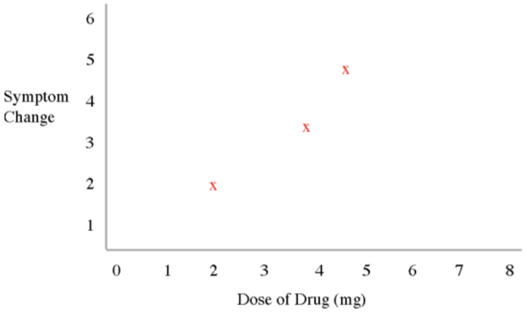


Figure 1. Data for response to three doses of a new drug

Imagine you are a clinical researcher and you've just done a study looking at three doses of a potential new medication: 2, 4, and 5mg. You want to know if you should study 7mg as a potential dose – will it help patients more than 5mg? and how much? You want to use a model to make a prediction about the response (change in symptoms) to a 7mg dose of medication. The simplest model you can use is a linear model so let's look at what the prediction in that case would be.

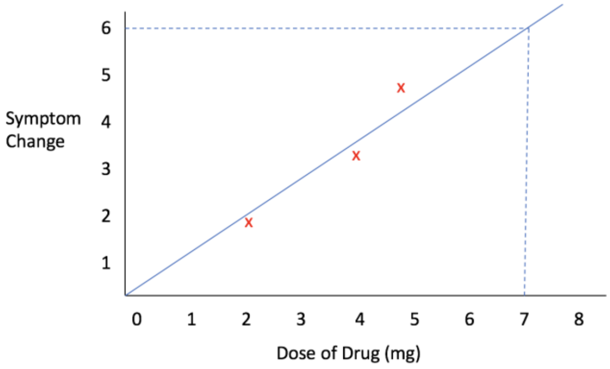


Figure 2. Data with a linear model (regression line) applied

We can see in figure 2, that the line (a linear regression, actually in this case a fake linear regression because I didn't actually run the data) shows that there should be an improvement in symptoms of 6 at a dose of 7mg. Based on this result you might predict that a dose of 7mg would be worthwhile to pursue for future research.

However if you go back to your pharmacology training you'll remember that dose response curves *don't* follow a straight line. Instead they follow a sigmoid curve because eventually the drug-target interaction saturates. Originally this was worked out by performing studies with many data points (probably at least 10) to flesh out a full curve. However, note that many pharmacology trials can serve as references, we no longer need to spend the time and money needed for this. Instead we can apply the right model from experience.

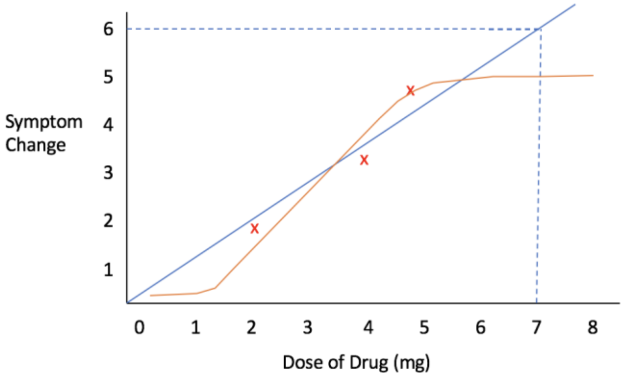


Figure 3. Data with linear and sigmoid model curves showing the difference in prediction

So in figure three we see the same data with a (poorly drawn) sigmoid model/curve fitted to the same data. With this model, we see that a dose of 7m is predicted to result in almost the same level of response as the 5mg dose. It's important to keep in mind that both of these curves are *predictions*, and in the absence of reliable data to be based upon, represent mathematical guesses. In other words, a good model is based on good knowledge. It is difficult for a non-statistician or investigator to evaluate the quality of the models used in an analysis, but if in doubt, knowing something about the authors (do they come from a good institution?) and the journal in which the paper is published (is it JAMA or one you've never heard of?) can help you feel confident someone with the right skill set evaluated the quality of analysis.

Pre-Guide

Diav-Citrin, O *et al* (2014). Pregnancy Outcome Following *In Utero* Exposure to Lithium: A Prospective, Comparative, Observational Study. *American Journal of Psychiatry* 171(7):785-794.

Reasons for choosing this article

- Understanding the risk and benefits of medications in pregnancy is an important clinical issue.
- This study lets us think about ways to answer questions other than randomized clinical trials, and the kinds of questions that are less amenable to RCTs.

Background

- What was your background knowledge about the safety of lithium in pregnancy?
- How would you design a study to investigate the effects of *in utero* lithium exposure? What do you make of the design the authors chose?
- What is the authors' hypothesis? What do you think of their choices of outcomes?

Methods

- How were pregnant women entered into the study?
- Which groups did the authors compare?
- How were pregnancy outcomes tracked? Whose report was used? How was it verified?

Results

- How many women were included in the study? How many years did it take to get this number of participants? What are your thoughts about this?
- What were the demographic and clinical differences between the groups? What are the implications of these baseline differences for the outcomes?
- What was the effect of lithium exposure of major anomalies? On cardiac anomalies?
- What do you make of the rate of pregnancy terminations in the lithium group compared to the other groups?
- How did the outcomes change with the addition of data from other sources? Why do you think the authors decided to add these data?
- What did the analysis of confounders with logistic regression add? How did the analysis for major anomalies and cardiac anomalies differ? What do these results mean?

Discussion

- What do you take away from this study?
- How would you counsel a woman with bipolar disorder about the risks and benefits of taking lithium during pregnancy?
- What are the implications for the way women were recruited for the study on the outcomes (consider selection bias)? What about the reliance on maternal report for outcomes?
- What do the authors mean by “detection bias?” Is this concern relevant to this study?

Post-Guide

Diav-Citrin, O *et al* (2014). Pregnancy Outcome Following *In Utero* Exposure to Lithium: A Prospective, Comparative, Observational Study. *American Journal of Psychiatry* 171(7):785-794.

Take Home Summary

This article describes the outcomes of prospective study of *in utero* lithium exposure among women who reported lithium exposure to the Israeli Teratology Information Service. Outcomes in women who took lithium during pregnancy were compared to women with bipolar disorder who did not take lithium and women counseled by the same service for nonteratogenic exposure. The authors found a higher rate of elective terminations and miscarriages among women treated with lithium as well as a higher rate of cardiac anomalies. There was no difference in the rate of major, non-cardiac anomalies. The higher rate of cardiac anomalies remained significant after controlling for several other factors and was observed among women treated with lithium in Australia and Canada. The strengths of this study are the prospective design and the relatively long follow-up for exposure outcomes. The dependence on maternal report for outcomes is a relative limitation. The inclusion only of women who chose to report lithium exposure to a national service is also a limitation, as these women may not be representative of all women exposed to lithium.

Pre-Guide

Donovan, NJ *et al*. 2016. Association of Higher Cortical Amyloid Burden with Loneliness in Cognitively Normal Adults. *JAMA Psychiatry* 73(12):1230-1237.

Reasons for choosing this article

- This study explores a critical area of geriatric psychiatry utilizing a cross-sectional design in a healthy population, which lets us think about the advantages and disadvantages of this study design.
- This study lets us think about two important things: (1) the difference between correlation and causation and (2) behavior as a biomarker.

Background

- The authors argue that identification of preclinical Alzheimer disease is important for secondary prevention. What do they mean by this?
- Why did the authors choose to study loneliness?
- What is the authors' overall hypothesis for their work? What is the hypothesis for this study?

Methods

- Who were the study participants? Were they recruited specifically for this study?
- How was loneliness assessed?
- How was amyloid level assessed?

A technical point from the Methods

What does it mean to utilize a "linear regression model" for data analysis? What does it mean to "control" for certain variables? How do you know what to control for?

Results

- How much loneliness was there in the study population? What are the implications of this?
- What relationship between amyloid and loneliness did the authors observe? Was this relationship consistent between the adjusted and unadjusted analyses?
- The authors perform an additional analysis comparing those defined as lonely to the other participants. How did they define lonely? What was the result of this analysis? Was this analysis consistent with the data presented in Figure 2?
- The authors perform an additional analysis comparing those defined as amyloid-positive to the other participants. How did they define amyloid-positive? What was the result of this analysis? Was this analysis consistent with the other analyses?

- Why do the authors perform a separate analysis of participants with a particular APO€4 status? Are the results here consistent?
- Why do you think they performed multiple analyses?

Discussion

- What do you take away from this study?
- Is this study important? Why or why not?
- The authors are very careful in their argument that they believe loneliness is a behavioral biomarker of amyloid. What do they mean by this? How is this different from saying that amyloid causes loneliness or that loneliness causes amyloid?
- What were the advantages/disadvantages of using a group of cognitively intact older adults? How might the results and interpretation have been affected if a different population (e.g, older adults with mild cognitive impairment, adults with suspected Alzheimer disease, or older adults with depression) was studied?
- Are there clinical implications of this work? If so, what are they?

Post-Guide

Donovan, NJ *et al*. 2016. Association of Higher Cortical Amyloid Burden with Loneliness in Cognitively Normal Adults. *JAMA Psychiatry* 73(12):1230-1237.

Take Home Summary

Alzheimer Disease (AD) is poised to weigh heavily on US society. Although no disease modifying therapies are currently available, there are a number of promising approaches on the horizon. Almost exclusively these therapies focus on delaying the clinical onset in people who have biological signs of disease, because it doesn't seem likely that brain damage from AD can be reversed. In order for this strategy to work, of course, we must develop effective ways of identifying people who are not yet patients and treat them years or decades before the onset of symptoms.

This article reflects one strategy: identify clinical symptoms, that, though subtle, predict very early disease development. Why not just use amyloid testing? The method used in this analysis, PET scanning, is the gold standard but expensive, technically difficult, and exposes patients to radiation, none of which are ideal for a screening test. Although it is likely that amyloid testing via PET will have an ongoing role in the clinical assessment for AD, it should not be the first step in screening. This analysis is part of that wider effort and builds on prior research that suggests that loneliness may be a relatively specific predictor. It is important to note that the authors propose that *feeling lonely* not *being alone* is a predictor of AD pathology.

Subjects in this analysis came from a larger study of healthy older adults with a primary aim of following the development of AD. A notable lack in the paper is detailed information on how and why the 79 subjects enrolled were chosen for inclusion in this sub-study. Subjects underwent PET scanning to detect amyloid and psychosocial assessment including a brief loneliness scale. The data were analyzed at a single time point to look for correlations between loneliness and the amount of amyloid present in the brain. They considered other factors that might influence amyloid accumulation in their models, discussed more in the technical point below.

The data showed a significant correlation between loneliness and the amount of amyloid present in the brain. This remains true as more possible covariates were considered, and if the group was split using a prior study's benchmark for "amyloid positivity." They also found that having the APOEe4 gene variant increased the strength of the association.

The authors discuss the finding primarily from the frame that loneliness is an early symptom of amyloidosis. Because their full model included anxiety and depression as cofactors, they felt confident that the feeling of loneliness was more specific than simply having depression; similarly, measures of social network quality also did not account for the association. There are two main caveats here. First, essentially all the subjects were below clinically relevant thresholds for these symptoms, and any time a sample is small and relatively homogeneous across a variable it may be hard to draw a conclusion accurately. This is especially important here as the connection between depression/anxiety and the development of dementia is fairly well replicated. Second, there is extensive data that support the idea that loneliness could cause inflammation and amyloidosis. It is likely that the relationship between the two is bi-directional, simply because it's hard to think of a biobehavioral phenomenon that isn't! Give that this is a cross-sectional study it really can't shed any light on causality. For the purpose the investigators engaged in this project however, causality doesn't much matter. Loneliness may still be a clinically relevant metric for risk-stratifying cognitively normal adults.

Technical Point

This paper includes three "linear regression models" examining the relationship between amyloid and loneliness. Why did the authors choose to analyze the data three times in different ways and report each of them in the article? First, let's talk about what linear regression is. At its most conceptually simple, linear regression is sort of like multivariate correlation – a regression of one variable will give the same result as a simple correlation. In theory, a regression allows you to consider the independent contribution of many variables to the outcome variable (in this case, amyloid deposition). Each of the variables generates a term that effects the slope of the "regression line." In general, the more terms included in a regression, the better the overall fit of the data, meaning that putting many variables in the model is a good way to 'hack' a positive result.

Because of this, it can be controversial how to choose variables for a model. Some teams only include variables that have previously been shown to be related to the outcome in question. So, here, age is clearly related to amyloid deposition so it should be included. Sometimes this is defined narrowly as variables correlated to the outcome in the study's own data set. The issue with this is that sometimes variables chosen by this method are materially not independent from the variables of interest (that is they are not true confounders). So, for example, depression has previously been associated with amyloid deposition. However, including it in the model may not be meaningful if depression and loneliness aren't really distinguishable phenomena. Of course, this method can also leave out important variables if they have no historical data or if the correlation has a meaningful value but a non-significant p-value. Because of this, some teams prefer to use their expert judgement about what to include rather than using a pre-defined rule.

In this paper, the authors may have tried to forestall criticism of the analysis by doing it both ways, essentially. Luckily for them both models showed significance. Often however this isn't the case, and significance can hinge on the exact combination of variables chosen. This is why best practice is to choose the list before doing analysis and to be as transparent as possible about the process followed. When you read papers, look for transparency in the methods. For big studies, a "design paper" is often published while the study is ongoing, and if it doesn't match up with what was done in the "outcome paper" that's a big red flag.

Pre-Guide

JE Dunsmoor *et al*. (2019) Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 39(17): 3264-3276.

Reasons for choosing this article

- This article is more focused on neuroscience than the articles we typically read, and you may find a bit difficult. That's okay! Do your best to get the big picture ideas.
- This article lets us revisit some of the foundational principles of associative learning thought to be related to the pathophysiology and treatment of anxiety disorders.
- This article is a really nice example of the type of human laboratory neuroscience research that gives rise to much of our understanding of how brains work.

Background

- Remind yourself of the definition of the following terms from classical (Pavlovian) conditioning: conditioned stimulus (CS), CS+, CS-, unconditioned stimulus (US), conditioned response (CR), unconditioned response (UR)
- What is extinction? Is extinction the same of as forgetting, or is it an active learning process?
- What do the authors mean by the term "associability?"
- In your own words, what are the authors trying to test in this study? What are their hypotheses?

Methods

- Who were the study participants?
- What were the steps in the study—what did the participants do first, then second, etc? (see also Figure 1A)
- What outcomes did the authors measure? When did they measure these things? What is skin conductance response (SCR)?
- What are the authors trying to assess with the presented equations? Don't worry about the math—focus on the theory and the goal/point of these equations.

A technical point:

- Here is a typical sentence for how the authors describe their findings: "Repeated-measures ANOVA using CS type (CS+, CS-) as a with-in subjects factor and group (EXT, NFE) as a between-subjects factor showed a main effect of CS type but not group and no CS type x group interaction." Explain the following terms: repeated-measures, between-subjects factor, with-subjects factor, main effect, interaction.

Results

- Figure 1 presents the behavioral results. Start with Figure 1B. What is on each axis? What are the different color bars? How do you interpret the information in this figure? What about Figure 1C?
  - Related to the above question, explain the following sentences from page 3267 that describe the results at the 24 hour recall test: "Independent samples *t* tests on the CS+ trials alone showed heightened mean SCRs in the EXT versus NFE group ([stats]) but no difference in the mean SCRs to the CS- ([stats])."-->This is the major behavioral finding of the paper.
- Now look at Figure 2. What brain regions responded differently during standard extinction vs novelty-facilitated extinction?
- What information is presented in Tables 1 and 2?
- The authors report the following about the whole brain analysis of retention test (page 3268): "Further analysis of the vmPFC confirmed that differences between CS+ and CS- in the EXT were driven by deactivations to the CS+ relative to the CS-, whereas the NFE showed relative increase in activity to the CS+ that was near the baseline level of the learned safety signal, the CS- (Fig. 3A)." What does this mean? Why is it important?

Discussion

- What do you take away from this study?
- What is the overall argument the authors make about the importance of novelty for learning?
- Why do you think we are reading this paper in journal club? How is it relevant to psychiatry?
- What type of study should be done next?

Post-Guide

JE Dunsmoor *et al*. (2019) Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction. *The Journal of Neuroscience* 39(17): 3264-3276.

Terms

Conditioned Stimulus (CS) – a neutral stimulus that normally doesn't provoke a strong response. If you think of the famous experiment of Pavlov's dogs, the conditioned stimulus was the bell ringing.

Unconditioned Stimulus (US) – a stimulus associated with innate positive or negative response. A shock, as given in this study causes a negative reaction (increased arousal, stress) without the need to learn it is unpleasant.

Conditioned/Unconditioned Response – the emotional or physiologic responses associated with events that are learned or unlearned, respectively. Sometimes unconditioned responses are, technically, learned, as long as they have become automatic within the context of the study. It's also possible to learn associations that counter the unconditioned response, as may happen with people who self-injure.

Extinction – the term for 'unlearning' a response to a conditioned stimulus. If Pavlov rings his bell without feeding his dogs every day for a month, the response will extinguish and they will no longer expect food when they hear the bell.

Take Home Summary

This is a directly translational study on fear memory that is relevant to PTSD and other anxiety disorders. Often anxiety pathology is pathology because a negative response occurs to things that 'shouldn't' provoke one. Often, as in PTSD, the person has learned 'too well' of danger and experiences the fear response even though they may know they are safe. In the language of this paper, that means that their conditioned response has not extinguished. Many experiments in animals have defined both the outward processes of conditioning as well as the neural basis of learned responses, and with the advent of fMRI, it became possible to study the neural functioning of human subjects undergoing conditioning and extinction paradigms as well.

The investigators studied a way to improve the extinguishing of negative conditioned responses in healthy adults. Their hypothesis was that teaching a person a new association for a CS would allow them to more fully extinguish a conditioned fear response. They used pictures of angry faces as the CS, and an electric shock as the US. On the first day of the experiment the subjects underwent training to associate the angry face picture with the shock. The response was determined by measuring the physiological response (measured via skin conductance) as well as the brain response. Immediately following the conditioning they were randomized and underwent training to extinguish the conditioning. In the control group, the training was 'standard extinction:' they were exposed to the picture without the shock. In the experimental group, the subjects were exposed to the CS and then instead of a shock they heard a tone. The next day they were brought back to the lab and underwent testing to see how well they had extinguished the original association between the faces and the shock.

The behavioral results (measuring skin conductance) showed that the group that learned a new association in the extinction training responded less to the CS compared to a dummy stimulus on the test day. The imaging results (which can seem overwhelming due to the large number of brain regions, focusing on figure 2 is suggested) supported that the experimental group was more active in several brain regions needed to form associations during the training. This group also showed more activity in regions thought to suppress the original association (the vmPFC and the superior frontal gyrus) during testing the next day. This was associated with less activity in other regions that would 'recall' the association. The authors conclude that simply keeping the learning regions more active by teaching a new association improved the learning to 'forget' the original conditioning. A very simple way of putting this is that the new stimulus made subjects 'pay more attention' to the extinction training. These results emphasize that extinction is an active learning process.

This same group of investigators is now looking at patients with anxiety disorders and PTSD to determine if their brains function more like the control group, and if using novel association paradigms can also help them extinguish associations. If so, new therapies could be developed to treat symptoms. Additionally, if the hypothesis that activity level in some brain regions improves extinction, then therapies that stimulate those brain regions, such as TMS, may be helpful for patients with anxiety disorders.

Technical point

Repeated Measures – the study involves the same data/measurement collected in the same people/animals/cells at multiple time points.

Between Subjects – the comparison of two or more groups in a study. You usually only hear this term in studies with repeated measures to refer to a comparison between groups at a single time point.

Within Subjects – the converse of *between subjects:* a comparison in a repeated measures study that look across time points in a single group.

Main Effect – a term used in regression analysis and other linear models to describe the contribution of one independent variable to the outcome, if you hold the other variables in the model 'stable' (which can be done as 'fixed' or 'random' effects).

Interaction – in contrast to the main effect, the way two independent variables, that are themselves related, together contribute to the outcome. For example, an inflammatory marker becomes much more important in predicting depression treatment outcome only as the severity of a medical comorbidity increases. You may not see much of a main effect for the marker alone but a big effect when accounting for the medical comorbidity as well, in their interaction.

Pre-Guide

NA Harrison et al. (2009). Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity.

Reasons for choosing this article

- The interaction of inflammation, the brain, and psychiatric symptoms (psychoneuroimmunology) is a very hot topic. This study is an introduction to this area, which you will likely hear more about in the future.
- The study subjects are healthy controls, and this study lets us think about when healthy people are the best people to study.
- This paper lets us think about the relevance of brain imaging lab studies to clinical care.

Background

- What is "sickness behavior?" What is the relevance of sickness behavior to depression?
- What patterns of mood symptoms are seen in people treated with immunotherapies like interferon-α?
- How might the immune system communicate with the brain?
- What hypothesis do the authors have for the study?

Methods

- Who were the study participants? Why did the authors recruit healthy volunteers to be the study participants? How many participants were there? Is this a big or small number of participants?
- What method was used to assess mood symptoms during the study? Why not use a questionnaire like the PHQ-9 that is used in clinical practice?
- What did the study participants actually do while they having the MRI done? (see Figure 1 and pg 409, paragraph starting "Twenty faces (10 male) from a standardized . . .)
- Briefly, what comparisons did the authors make? E.g., Did they compare the same patient on typhoid vaccine vs placebo? Did they compare how the subjects responded to see happy vs sad faces? Did they look at the whole brain? Some selected brain areas? Some combination of these things?

A technical points from the methods:

- In imaging studies, authors often conduct a "whole brain analysis" and a "region of interest analysis." What do these terms refer to?

Results

- What effects did typhoid vaccination have on the immune system of the study participants? What effects did vaccination have on mood? Were these associated with each other?
- Figure 2 describes a relationship between mood symptoms and detection of MRI brain signal. In your own words, describe this figure. What brain region was affected? In what way has this region been implicated in depression?
- Were the changes restricted to the area described in Figure 2? Or were areas associated with this region also affected?

Discussion

- What do you take away from this study?
- Are there implications of this study for clinical practice? If so, what are they? If not, why not?
- Some people have started to argue that depression is a disease of the immune system. What do you think of this argument?
- What might be some important next steps in research based on this study?

Post-Guide

NA Harrison et al. (2009). Inflammation Causes Mood Changes Through Alterations in Subgenual Cingulate Activity and Mesolimbic Connectivity.

Summary

We chose this paper to show how scientists use "translational neuroscience" techniques to explore basic questions of human neurophysiology and to highlight the field of psychoneuroimmunology (PNI). PNI examines the interplay between immunity and neural function. Immune system abnormalities have been found in almost every major mental disorder from anorexia to schizophrenia, and mood disorders have the best evidence that immune activity is a routine causative factor. The background of this paper does a great job of providing a quick review of the evidence that exogenous inflammation is causally related to the development of depression for at least some patients.

The purpose of this study was to learn more about the effect of inflammation on specific brain circuits that regulate mood. Although this model is not exactly a model of depression – it is normal to feel sick after receiving an antigen such as a vaccine – it still provides insight into how the brain receives and responds to inflammation. Typhoid vaccine is composed of dead *Salmonella typhi* bacteria and provokes a strong but brief response from the immune system. Using fMRI, this research group examined how the brain's response to viewing emotional faces changes during this time of the immune response to the vaccination. Like many translational studies, this one used a small sample of healthy young men. The participants were screened to rule out background inflammation and enrolled into a crossover designed protocol. This means that each subject served as his own control by coming twice and undergoing the same procedures except that at one visit a vaccine was given, and at the other placebo was given. Some participants received the active vaccine in the first session, and others received placebo first ("counterbalanced"), which helps ensure that differences between conditions are not related to effects of experience with the task.

Three hours after vaccination, participants were put in an MRI scanner and were shown a series of photographs of people with emotional expressions. The subjects were tested on *implicit* emotional processing: they were instructed to guess the ages of people in photographs rather than to think about or act on the facial expressions. The primary analysis of the data divided the runs according to emotions displayed. There is a fairly extensive literature that has demonstrated stronger implicit responses to negative emotional expressions than positive emotions in patients with depression, so this task was chosen to be comparable to with prior findings.

Three hours after receiving the typhoid vaccine, participants reported significantly worse mood than prior to vaccination and had higher levels of IL-6, an inflammatory cytokine; these changes were not seen with placebo vaccination. In the placebo condition, activations in brain regions usually associated with seeing human faces and processing emotions were found. When comparing active vaccine to placebo, more activity in the subgenual anterior cingulate cortex (sACC) was seen, while there was less activity in the amygdala. The authors performed an analysis to trace back the connectivity between these regions and other areas of the brain and found that worse mood was associated with the weaker connection between the medial prefrontal cortex (MPFC) (specifically the anterior rostral portion) and the sACC.

Other studies have consistently demonstrated an association between increased sACC activity and depression. The sACC appears to integrate self perception with our understanding of the world and the social hierarchy. Over activation in this region in depression seems to be related to failure of the frontal cortex to regulate sACC activity, which is similar to the findings in this study. We also know the amygdala is important in signaling salience – that is, what you should be paying attention to – and it fails to perform this job properly in depressed patients.

Overall this paper demonstrates that the effects of inflammation on the brain acutely resemble chronic changes in brain activity present in depression. The biggest remaining question is what factors may lead to the shift to a chronic state and whether inflammation interacts additively with stress and other depression risk factors to cause long term changes in the brain.

Technical Point

Brain imaging has a number of analytic design choices that must be made when putting together a study. In this case because the authors began the project with specific brain regions in mind as relevant to regulation of mood and for viewing faces; they knew fairly well where to look in the brain for differences of interest. Such a list of brain areas are typically called “regions of interest” (ROIs). In other words, they had a specific hypothesis to test about how inflammation effects the brain. In the absence of this, investigators may use “whole brain” techniques that don’t exclude any part of the brain, but can be prone to false positive results. Whole brain analysis is most appropriate for hypothesis generating studies. In both cases, a task that is likely to cause activity in relevant brain regions must be used. An obvious case is that if you are studying vision you shouldn’t use a task involving sound, though most functional imaging research involves a need to match the task to the brain function much more subtly. For example, in this case the sACC isn’t a region usually associated with face processing but was nonetheless associated with the task here. The authors chose this region for analysis based on prior reports of depression-related brain circuitry. Interestingly, this difference in activity was seen when comparing trials of emotional to neutral faces and examining the differences in brain activity associated with worse mood during the scan. A sharp critic of this paper may note that these results have limited interpretability since there was a task-ROI to syndrome-ROI mismatch. Although it’s very difficult to understand the nuances (or even sometimes the broad strokes) of imaging study design and analysis, you should be able to keep in mind – does this study test a specific hypothesis with specific (ROI-based) methods? Or does it ask a general question using more open ended whole brain methods? In both cases, do the authors present evidence that the task the participants complete relates convincingly to the symptom or illness under study? (Thanks to Dr. Carrie McAdams for providing critique of the methods in this paper.)

Pre-Guide

W. C. LaFrance Jr *et al*. (2014). Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 71(9): 997-1005.

Reasons for choosing this article

- Psychogenic nonepileptic seizures are a common reason for psych consults, and it's a condition that many residents see during neurology rotations.
- This paper describes a small pilot trial that was published in a high impact journal, and it's interesting to consider why that might be.
- This article allows us to again consider efficacy of psychotherapy compared to medication, a common clinical question.

Background

- What has been your experience with patients with PNES? How was the diagnosis made and discussed with the patient? The authors note that many patients diagnosed with PNES do not seek mental health care--does that fit with your experience?
- Why do the authors use the term "psychogenic nonepileptic seizures" and not "pseudoseizures"?
- Prior to this study, what was known about psychotherapy in the treatment of PNES? What about the role of sertraline? What was the rationale for choosing to test each of these interventions?
- What do you think the hypothesis for the study was?

Methods

- The study recruited 38 patients over a 3.5 year period from 3 sites. What do these numbers tell you about rate of recruitment? What do you think explains this?
- What do you think about the choice to have seizure frequency as the primary outcome?
- What kind of training did the psychotherapists receive? Is this training a strength or limitation of the study?
- What is a blinded rater? Why were blinded raters used?

A technical point from the results:

- On page 1001, the authors state: "The pilot study was not powered to detect between-group differences and was designed for within-group analyses." What does this mean? What are "between-group differences" vs "within-group analyses"?

Results

- The authors give information on "screen failures." What is a screen failure? Why do they report this? (See also Figure 1)
- Looking at Table 1--what do you notice about the rates of comorbidities, both overall and between treatment groups? What were the common comorbidities (and how common were they)?
- What was the effect of each intervention on seizure frequency? Were there effects on any of the secondary outcomes?
- On page 1001, the authors report what the patients in each arm expected the outcome to be. Why did they ask the trial participants about their expectations?
- The authors note that baseline scores on measures of anxiety and depression differed between groups. Does this cause problems with interpreting the results? How do the authors attempt to address this in their data analysis? Does their method seem sufficient to you?

Discussion

- What do you take away from this study?
- The authors provide a theory for the partial efficacy of sertraline alone--do you agree with their theory? Is this something that could be tested?
- Do you think CBT-ip as described in the current study could be widely disseminated so that the many patients with PNES would have access to this treatment?

Post-Guide

W. C. LaFrance Jr *et al*. (2014). Multicenter Pilot Treatment Trial for Psychogenic Nonepileptic Seizures: A Randomized Clinical Trial. *JAMA Psychiatry* 71(9): 997-1005.

Take Home Summary

Psychogenic non-epileptic seizures (PNES) is a relatively common diagnosis for psychiatrists consulting on neurological patients; as this article points out a sizable fraction of work-ups in epilepsy monitoring units result in this diagnosis rather than 'true' seizures (and many patients with epilepsy have both types, to complicate things further). Despite the significant functional impairment associated with PNES, evidence- based treatment is almost totally lacking. This article describes the results of a small pilot trial of four interventions for PNES: treatment-as-usual, sertraline alone, cognitive behavioral therapy-informed psychotherapy (CBT-ip), or CBT-ip with sertraline. Although this study was quite small the design is high quality and the content addressed a significant gap in literature. The impact of the project was high enough to make publication in a fairly high impact journal worthwhile. It probably also helped that this is an illness that crosses the paths of more than one specialty. The most common treatments for PNES are medications (SSRIs which are of course used for depression and anxiety, heavily comorbid with PNES) and psychotherapy, so these are the interventions tested here. The study participants were adults who had documentation that excluded epileptic seizures. While patients with substance use disorders, psychosis, or self-harm were excluded, patients with other psychiatric comorbidities were included.

The trial lasted a total of 16 weeks, with the number of episodes being the primary outcome. During the first two weeks, subjects were monitored to get an accurate baseline seizure frequency and then treatment was started at week 2 and continued throughout the trial. Sertraline was titrated to 200 mg as tolerated; psychotherapy, which was specially designed for PNES patients, was administered in 12 weekly 1-hour sessions. Participants in the treatment as usual group were seen for assessments on the same schedule as the other participants. The authors screened 589 patients, of whom only 81 were eligible. More interestingly, only 38 agreed to sign consent and three of those immediately changed their minds, with another withdrawing shortly thereafter.

At the end of the trial, decreases in seizure frequency by group were as follows: CBT-ip 51.4%, CBT-ip with sertraline 59.3%, sertraline alone 26.5%, and treatment as usual 33.8%. For CBT-ip and CBT-ip with sertraline, these decreases were significant compared to baseline; the changes were not significant in the sertraline alone or treatment as usual groups. The odds of being seizure free were 6.2 times greater for patients receiving CBT-ip compared to those not receiving this treatment, and patients receiving CBT-ip were less likely to be seen in the ED. Patients receiving CBT-ip plus sertraline also had decreases in measures of anxiety and depression. This trial presents preliminary evidence for the efficacy of CBT-ip in the treatment of PNES. It also presents preliminary data that sertraline is effective for the treatment of comorbidities associated with PNES but does not have an effect on seizure occurrence.

Technical point

The authors discuss performing two different types of comparisons: ‘within group’ and ‘between group.’ Within group comparisons describe changes within a treatment group--the rate of seizures at baseline compared to seizure rate after 12 weeks of treatment with CBT-ip. Between group differences are comparisons across treatment groups--rate of seizures at the end of the trial in the CBT-ip group compared to the sertraline group. In designing this small trial, the authors, probably knowing that recruitment would be difficult and lacking evidence that any of the treatments would be sure to work, chose to power the study such that they would be able to detect efficacy of each treatment arm alone, but not determine which, if any, was superior. Thus, while the within group comparisons demonstrated that CBT-ip and CBT-ip with sertraline decreased seizure occurrence significantly, they didn’t directly compare these two treatment arms. This design is appropriate for a pilot study, in which demonstrating that the intervention has an effect and measuring its size are appropriate goals, saving a more accurate assessment of treatments relative to on another for a larger trial. In particular, it may help a future trial by justifying not including a treatment as usual arm, allowing more subjects to be randomized to viable treatment options.

Pre-Guide

JD Lee *et al*. (2018) Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 391(10118): 309-318.

Reasons for choosing this article

- This manuscript describes the primary outcomes of the X:BOT trial, which addresses a major question in addiction psychiatry related to the ongoing opioid epidemic.
- The study design and analysis demonstrate the push and pull between rigorous science and pragmatism.

Background

- What are the major differences between treatment with naltrexone and buprenorphine for opioid use disorder? Why are there concerns about safety with naltrexone?
- What do the authors give as the reasons for conducting this study?
- What was the hypothesis of study?

Methods

- Who were the study participants (i.e., what were the inclusion and exclusion criteria)? How were they recruited for study participation?
- To what extent did the study control the detox phase of treatment? What do you think of this? To what extent was buprenorphine dosing regulated by the study?
- What was the primary outcome? How was it defined? What do you think of this definition? At what point in the study was the primary outcome assessed?

A technical point from the Methods:

The authors perform two types of analyses—the first is the intention to treat analysis, and the second is the per protocol analysis. What is meant by these terms? What are the differences between the analyses?

Results

- What are the major findings of the study? What differences were observed between the treatment groups? How did the type of analysis effect the results? In other words, describe Figure 2.
- What is meant by the “induction hurdle”? How did this effect the outcomes?

- Were there differences between groups in adverse events or in deaths?

Discussion

- What do you take away from this study?
- The authors list “five major findings” from the study. What are they? Do you agree with their list?
- Some researchers have criticized the use of the per protocol analysis. What do you think the criticisms are? Do you think they are valid?
- Toward the end of the discussion, the authors raise the issue of study retention. How was retention in the study? Are there statistical concerns when study retention is low?

Post-Guide

JD Lee *et al*. (2018) Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial . *Lancet* 391(10118): 309-318.

Take Home Summary

This study addresses a pressing clinical issue in the US today – how best to treat patients who are addicted to opioids to prevent relapse and overdose. Specifically, it compares two newer long term therapies – long acting naltrexone (XR-NTX, an injection) and buprenorphine-naloxone (BUP-NX, in this study given as a sublingual film) given over 24 weeks. Given the scope of opioid addiction and the large number of deaths, often focused in areas with low financial resources, showing the feasibility and outcomes and economic efficiency of opioid maintenance treatments is critical to bring these therapies to the largest number of patients.

Subjects were adults who presented for opioid detox and were enrolled in the study once they were admitted to one of eight detox sites. Because the process of initiation of the two drugs differs, there was flexibility in how subjects were enrolled, when they were randomized, and how long they spent at the sites before starting study drug. Once started, they were maintained, non-blinded, on study drug for up to 24 weeks with some post treatment follow-up. The primary outcome was relapse of use which, is as is often the case in these studies, has a complicated definition involving both urine drug screens and self-reported use. As with most studies of addiction, subjects who drop out without any information available about their outcome are assumed to have relapsed. A long list of secondary outcomes were examined, including the rate of successful induction on therapy, adverse events including overdose, and self-reported opioid craving, were also assessed.

Overall the study found both treatments were equally effective in preventing relapse of opioid use, though these rates were only about 50%. However, this finding applies only to those who were successfully inducted onto therapy (discussed further below). These findings applied not only to relapse prevention but also to all the secondary outcomes, supporting that overall these two therapies are both valuable options in the treatment of opioid use. Most of the overdose events in the study occurred in those who were not on medication, and although the study was not powered to look at this statistically, it should suggest that prevention of death by overdose is an important part of medication treatment for opioid addiction.

Technical Point

This study attempted to study real world effectiveness of opioid relapse prevention therapies in a very messy real world. As the authors noted in the discussion, there is wide variability in the treatment of opioid use. The eight sites participating in the study varied in ways the authors knew would affect the outcomes, and the two treatments being compared have significant procedural differences in their prescription/induction (how the medication is started). In particular XR-NTX requires that a person be fully detoxed prior to induction, which means that it takes longer from presentation. This caused significant variation in the ability of sites to successfully start subjects randomized to XR-NTX on therapy, based on the length of stay, the detox methods (e.g., some of the study sites used buprenorphine as the detox treatment), and the level of support subjects received while awaiting induction. Even without considering the differences between the treatment arms, there is substantial variability across subjects in their motivation to stop using and their ability to successfully complete detoxification. For this reason, the investigators tracked when, relative to presentation, the subjects were randomized. Early randomizers had not completed detox and had to wait before starting study drug, especially if it was XR-NTX. Late randomizers had more fully completed detox when enrolled. It's possible that if drop-out during detox is high, subjects in the late randomization group also differed statistically in other subtle or non-measured ways – such as psychosocial stressors and motivation for sobriety – than the early randomization group.

These differences led to the analysis in which only those who successfully inducted onto medication were analyzed side by side with the analysis in which all enrolled subjects were analyzed. The differences between the two add substantially to the overall impact of this paper and help suggest directions for future research in this area. First the induction hurdle for XR-NTX is substantial. However once crossed treatment is equally effective. This may be important for those who don't want to take an opioid like BUP-NX, who would prefer to not to take daily medication, or for other reasons prefer XR-NTX.

From the perspective of learning about clinical science, it's important to note that there is no single objective standard way to analyze study data. In this case using two parallel methods highlighted patterns in the data that neither analysis method alone could have.

For more information on how the authors approached key issues in the study design, see: EV Nunes *et al*. (2016) Ethical and clinical safety considerations in the design of an effectiveness trial: a comparison of buprenorphine versus naltrexone treatment for opioid dependence. *Contemporary Clinical Trials* 51:34-43.

For a brief overview on intention-to-treat analyes, see: MA Detry and RJ Lewis. (2014) The intention-to-treat principle: How to assess the true effect of choosing a medical treatment. *JAMA* 312(1):85-86.

For a more detailed overview of per-protocol analyses, see: MA Hernan and JM Robins. (2017) Per-Protocol Analyses of Pragmatic Trials. *NEJM* 377(14): 1391-1398.

Pre-Guide (PGY2-4)

Lieberman JA *et al*. 2005. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine* 353(12):1209-1223.

Reasons for choosing this article

- This article reports the primary outcomes of phase I of the CATIE trial, a major effectiveness trial for antipsychotics in schizophrenia. It is a landmark study.
- The CATIE methodology has been heavily criticized, with some arguing that the trial results are invalid due to problems with the methodology.

Background

- Prior to this study, what was thought about the differences between typical and atypical antipsychotics? The authors state that atypicals were argued to have “enhanced safety and efficacy”—what is this referring to?
- What were the reasons for doing this trial?
- What was the authors' hypothesis?

Methods

- What do you think was the rationale for making the protocol available for comment prior to completing the study?
- What do you make of the number of sites in the study? Is this typical?
- How were patients randomized to treatments? In which circumstances was a patient prevented from being randomized to a certain treatment? What are the implications of this?
- Look closely at the section describing the medication dosing, as the way the dosing was done was one of the largest critiques of CATIE. Consider both the dosing as described in the methods and the paragraph on page 1212 that provides the mean modal doses used in the study.
- Why did the authors choose discontinuation of treatment as the primary outcome? Do you agree with this choice?
- Do you think the study had appropriate power?

A technical point from the Methods:

What is an “intention-to-treat” analysis?

Results

- What are the results with regards to the primary outcome (discontinuation for any cause)? Did any drug stand out? Did this change based on the outcome measure (i.e., discontinuation for any cause vs discontinuation due to lack of efficacy vs discontinuation due to side effects)? Hint: Figure 2 (which is easiest to read in color) presents this information succinctly.
- Looking at Table 1: how do the patients in the trial compare with the patients with schizophrenia you've seen in residency? Do the trial participants seem to be representative of patients with schizophrenia?
- Overall, how long did patients remain on their assigned treatment?
- How many patients gained weight on olanzapine? How much did they gain?

Discussion

- What do you take away from this study?
- Some people have argued that the CATIE trial demonstrates that there is no difference in efficacy between typical and atypical antipsychotics. Is this a fair summary?
- Do you think the results of CATIE are related to the choice of perphenazine as the typical antipsychotic? Do you think the results would have been different if another agent (haloperidol, thorazine) had been used?
- The authors state that their results "might lead one to consider olanzapine the most effective of the medications." Do you agree? Are there any limitations to this statement?

Much has been written about the CATIE trial. For commentaries on this trial, see:

- JA Lieberman and TS Stroup. The NIMH-CATIE Schizophrenia Study: What Did We Learn? *AJP* 108(8): 770-775.
- S Lewis and JA Lieberman. 2008. CATIE and CUtLASS: Can We Handle the Truth? *British Journal of Psychiatry*. 192(3):161-163.
- H. Moller. 2005. Are the New Antipsychotics No Better Than the Classic Neuroleptics? *Cur Arch Psychiatry Clin Neurosci* 255:371-372.
- All of the articles in the May 2008 issue (volume 59, issue 5) of *Psychiatric Services* are commentaries on CATIE.

Post-Guide (PGY2-4)

Lieberman JA *et al*. 2005. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine* 353(12):1209-1223.

Article Summary

This manuscript reports on the main outcome results of a large effectiveness trial of antipsychotic medications in patients with schizophrenia. **The primary measure of effectiveness was discontinuation of the study medication for any reason.** Subjects were adults with an established diagnosis of schizophrenia – patients with related psychotic disorders were excluded but most other psychiatric and stable medication conditions were allowed. Medications used in the study were perphenazine, risperidone, quetiapine, olanzapine and, later, ziprasidone. Patients were pseudo-randomized to one of these drugs in a double blind fashion, and were able to take from 1-4 pills a day for the "best" dose balancing efficacy and side effects. Subjects were followed for up to 18 months. A Kaplan-Meier survival curve analysis was used to compare the time to discontinuation for the 5 drugs.

The study found that the majority of subjects in all arms discontinued treatment (74%). When comparing discontinuation by group, olanzapine had the lowest absolute discontinuation rate at 64%. Because of the addition ziprasidone midway into the trial, the between group analysis was done "without" and then "with" ziprasidone. **Comparing the original four groups, olanzapine was superior to quetiapine and risperidone but not perphenazine, once correction was done for multiple testing**. When the ziprasidone group was added to the analysis, it did not differ significantly from the other treatments after correction for multiple testing. When examining discontinuation for poor efficacy, the pattern was similar, with olanzapine slightly out-performing other drugs, but with significance often disappearing when correcting for multiple testing. No differences between drugs was found for discontinuation due to side effects.

Comments

The CATIE trial is the largest and most significant study to address the issue of whether there are real meaningful outcome differences among second generation antipsychotics and, between first and second generation agents. Pharmaceutical companies have no incentive to perform this type of trial, and are not required to do so by the FDA, so data of this kind is scarce. Following patients of any kind, but especially those with serious mental illness, for 18 months is a difficult and expensive task, making this study more than valuable enough for publication in a flagship journal like The New England Journal of Medicine. The main outcome – simply how long patients remain in treatment – is paradigmatic of **effectiveness trials**, which attempt to weigh all the factors effecting treatment in the "real world" and not just the technical effectiveness of a drug. In psychotic disorders, where adherence to treatment is a serious issue, this type of trial is even more important.

Overall, they found little difference between agents using the outcome of time in treatment. Of note, the addition of ziprasidone relatively late in the trial substantially reduced the **statistical power**, and was a major criticism of the study. One important finding is that although olanzapine was superior to the other second generation drugs, it was not superior to perphenazine. This is widely perceived as undermining claims by drug companies that second generation agents are, per se, better than older drugs.

Technical Point

An interesting feature of CATIE is the choice of “time on [study] treatment” as the primary outcome. Historically, studies often analyzed only those subjects who completed the study, but this eventually was identified as a source of error or bias in reported results because drop-out isn’t random. For example, if the treatment (or in CATIE, one of the treatments) causes significant side effects, more subjects drop out of that arm than the placebo or alternate treatment arm(s). To account for this, most trials today perform **“intent to treat”** analyses. This means that once a subject has taken any medication in a trial, his or her data must be included in the results. The term “intent to treat” may mean any of several different methods are used in practice to keep track of drop-outs. In CATIE, they chose the relatively un-orthodox but exciting method of considering drop-out (as well as other medication discontinuation) as an outcome. This is one important element in defining CATIE as an **effectiveness** trial.

Pre-Guide (Intern)

Lieberman JA *et al*. 2005. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine* 353(12):1209-1223.

Accompanying Design and Statistics Article

Kaji AH and Lewis RL. 2015. Noninferiority Trials: Is a New Treatment Almost as Effective as Another? *JAMA* 313 (23): 2371-2372.

Reasons for choosing this article

- This article reports the primary outcomes of phase I of the CATIE trial, a major effectiveness trial for antipsychotics in schizophrenia. It is a landmark study.
- The CATIE methodology has been heavily criticized, with some arguing that the trial results are invalid due to problems with the methodology.

Background

- Prior to this study, what was thought about the differences between typical and atypical antipsychotics? The authors state that atypicals were argued to have “enhanced safety and efficacy”—what is this referring to?
- What were the reasons for doing this trial?
- What was the authors' hypothesis?

Methods

- What do you think was the rationale for making the protocol available for comment prior to completing the study?
- What do you make of the number of sites in the study? Is this typical?
- How were patients randomized to treatments? In which circumstances was a patient prevented from being randomized to a certain treatment? What are the implications of this? Compare/contrast this to the randomization scheme in STAR-D.
- Look closely at the section describing the medication dosing, as the way the dosing was done was one of the largest critiques of CATIE. Consider both the dosing as described in the methods and the paragraph on page 1212 that provides the mean modal doses used in the study.
- Why did the authors choose discontinuation of treatment as the primary outcome? Do you agree with this choice?
- Do you think the study had appropriate power?

A technical point from the Methods:

What is an "intention-to-treat" analysis? The article on non-inferiority designs contrasts an intent-to-treat analysis against a per protocol analysis. What are the differences?

Results

- What are the results with regards to the primary outcome (discontinuation for any cause)? Did any drug stand out? Did this change based on the outcome measure (i.e., discontinuation for any cause vs discontinuation due to lack of efficacy vs discontinuation due to side effects)? Hint: Figure 2 (which is easiest to read in color) presents this information succinctly.
- Looking at Table 1: how do the patients in the trial compare with the patients with schizophrenia you've seen in medical school and residency? Do the trial participants seem to be representative of patients with schizophrenia?
- Overall, how long did patients remain on their assigned treatment?
- How many patients gained weight on olanzapine? How much did they gain?

Discussion

- What do you take away from this study?
- Some people have argued that the CATIE trial demonstrates that there is no difference in efficacy between typical and atypical antipsychotics. Is this a fair summary? Consider the article on non-inferiority trials in thinking about this.
- Do you think the results of CATIE are related to the choice of perphenazine as the typical antipsychotic? Do you think the results would have been different if another agent (haloperidol, thorazine) had been used?
- The authors state that their results "might lead one to consider olanzapine the most effective of the medications." Do you agree? Are there any limitations to this statement?

Much has been written about the CATIE trial. For commentaries on this trial, see:

- JA Lieberman and TS Stroup. The NIMH-CATIE Schizophrenia Study: What Did We Learn? *AJP* 108(8): 770-775.
- S Lewis and JA Lieberman. 2008. CATIE and CUtLASS: Can We Handle the Truth? *British Journal of Psychiatry*. 192(3):161-163.
- H. Moller. 2005. Are the New Antipsychotics No Better Than the Classic Neuroleptics? *Cur Arch Psychiatry Clin Neurosci* 255:371-372.
- All of the articles in the May 2008 issue (volume 59, issue 5) of *Psychiatric Services* are commentaries on CATIE.

Post-Guide (Intern)

Lieberman JA *et al*. 2005. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. *New England Journal of Medicine* 353(12):1209-1223.

Accompanying Design and Statistics Article

Kaji AH and Lewis RL. 2015. Noninferiority Trials: Is a New Treatment Almost as Effective as Another? *JAMA* 313 (23): 2371-2372.

Article Summary

This manuscript reports the main outcome results of a large effectiveness trial of antipsychotic medications in patients with schizophrenia. **The primary measure of effectiveness was discontinuation of the study medication for any reason.** Participants were adults with an established diagnosis of schizophrenia – patients with related psychotic disorders were excluded but most other psychiatric and stable medical conditions were allowed. Medications used in the study were perphenazine, risperidone, quetiapine, olanzapine and, later, ziprasidone. Patients were pseudo-randomized to one of these drugs in a double blind fashion, and medication was dosed as 1-4 pills a day for the "best" dose balancing efficacy and side effects. Subjects were followed for up to 18 months. A Kaplan-Meier survival curve analysis was used to compare the time to discontinuation for the 5 drugs.

The study found that the majority of participants in all arms discontinued treatment (74%). When comparing discontinuation by group, olanzapine had the lowest absolute discontinuation rate at 64%. Because of the addition ziprasidone midway into the trial, the between group analysis was done "without" and then "with" ziprasidone. **Comparing the original four groups, olanzapine was superior to quetiapine and risperidone but not perphenazine, once correction was done for multiple testing**. When the ziprasidone group was added to the analysis, it did not differ significantly from the other treatments after correction for multiple testing. When examining discontinuation for poor efficacy, the pattern was similar, with olanzapine slightly out-performing other drugs, but with significance often disappearing when correcting for multiple testing. No differences between drugs was found for discontinuation due to side effects.

Comments

The CATIE trial is the largest and most significant study to address the issue of whether there are real meaningful outcome differences among second generation antipsychotics and between first and second generation agents. Pharmaceutical companies have no incentive to perform this type of trial, and are not required to do so by the FDA, so data of this kind is scarce. Following patients of any kind, but especially those with serious mental illness, for 18 months is a difficult and expensive task, making this study more than valuable enough for publication in a flagship journal like The New England Journal of Medicine. The main outcome – simply how long patients remain in treatment – is paradigmatic of

effectiveness trials, which attempt to weigh all the factors effecting treatment in the "real world" and not just the efficacy of a medication compared to placebo. In psychotic disorders, where adherence to treatment is a serious issue, this type of trial is even more important.

Overall, they found little difference between agents using the outcome of time in treatment. Of note, the addition of ziprasidone relatively late in the trial substantially reduced the **statistical power**, and was a major criticism of the study. One important finding is that although olanzapine was superior to the other second generation drugs, it was not superior to perphenazine. This is widely perceived as undermining claims by drug companies that second generation agents are, per se, better than older drugs.

Technical Point

An interesting feature of CATIE is the choice of "time on [study] treatment" as the primary outcome. Historically, studies often analyzed only those subjects who completed the study, but this eventually was identified as a source of error or bias in reported results because drop-out isn't random. For example, if the treatment (or in CATIE, one of the treatments) causes significant side effects, more subjects drop out of that arm than the placebo or alternate treatment arm(s). To account for this, most trials today perform **"intent to treat"** analyses. This means that once a subject has taken any medication in a trial, his or her data must be included in the results. The term "intent to treat" may mean any of several different methods are used in practice to keep track of drop-outs. In CATIE, they chose the relatively un-orthodox but exciting method of considering drop-out (as well as other medication discontinuation) as an outcome. This is one important element in defining CATIE as an **effectiveness** trial.

The accompanying statistics article contrasts an intent-to-treat analysis against a per protocol analysis. In a per protocol analysis, only data from the participants who complete the study as it was designed (i.e., complete the study per the study protocol) are included in the analysis.

Accompanying Design and Statistics Article

This article gives a description of the design and goals of a non-inferiority trial, which is utilized to determine that a given treatment is no worse than another. A non-inferiority design is different from a traditional trial, which seeks to determine that a treatment is different (typically better) than control, and it is different from an equivalence trial, which seeks to determine that two interventions are equivalent to each other. In designing and performing a non-inferiority study, the study investigators set the definition of "no worse" (the non-inferiority margin), which is often subjective. Because a non-inferiority trial is seeking only to determine that the new treatment is no worse, a one-sided *p* is appropriate. The CATIE trial is not a non-inferiority trial. This stats article was paired with CATIE to highlight how CATIE differs from a non-inferiority trial. CATIE sought to address a question that could have been been appropriate for a non-inferiority design. The authors could have asked the question as "is a typical antipsychotic no worse than an atypical antipsychotic." Instead, they (in effect) posed the question "are any of the tested antipsychotics better than the others."

Pre-Guide

M.M. Linehan *et al*. (2015). Dialectical Behavior Therapy for High Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 72(5): 475-482.

Reasons for choosing this article

- The article allows for discussion of DBT, a topic which many residents would like to learn more about.
- This article provides the opportunity to read and discuss a clinical trial of a therapy modality.
- It lets us think about the ways in which therapy may be implemented in ways different than the way it is tested and what it means to say that a particular modality is evidence-based.

Background

- From your experience, how easy is it for patients to get access to DBT? Why do you think this is the case?
- According to the authors, what are the reasons to perform this study?
- What is the hypothesis? Does this hypothesis seem reasonable?

Methods

- Does this study contain a control group?
- Do you think the groups used allow for good comparisons? Looking at Table 1—what are the differences between standard DBT, DBT-S, and DBT-I?
- Study participants were required to have a diagnosis of borderline personality disorder and recent self-injury with a history of suicide attempt or self-injury—why do you think that a history suicide attempt/self injury was part of the inclusion criteria?
- In describing the DBT individual therapy (p 477), the authors state: "to control for treatment dose, an activity-based support group was added . . ." What do they mean by "treatment dose?" What was the purpose of this support group?
- What do you make of the prohibition of teaching of DBT skills in individual therapy?

A technical point from the Methods:

The authors state they used an “adaptive randomization procedure” that “matched participants on age, number of suicide attempts, number of NNSI episodes, psychiatric hospitalizations in the past year, and depression severity.” What term is typically used to describe a randomization scheme in which the groups are created to be equivalent on certain variables? How is this different from “pure” randomization? How is this different from controlling for these variables in the analysis?

Results

- The authors chose incidents of self-injury as the primary outcome for the study. Why do you think they chose this outcome? What do you think are the appropriate outcomes to assess?
- What were the major findings of the study? On which outcomes did the groups differ?
- Do the authors make the appropriate group comparisons?
- What do you make of the difference in time to drop-out between standard DBT and DBT-I?

Discussion

- What do you take away from this study?
- At the beginning of the discussion, the authors state: “the focus of this randomized clinical trial was to determine whether the skills training component of DBT is necessary and/or sufficient to reduce suicidal behaviors . . .” How does this compare with the hypothesis in the Introduction?
- Related to the above, what are the “necessary and/or sufficient” components of DBT?
- In discussing the limitations of the study, the authors state “we were not willing to let someone die by suicide to make a point.” The authors go on to argue that the procedures used in the study to minimize the occurrence of suicide also limited their ability to detect differences between groups. Specifically, what are they referring to? Do you agree that doing things differently would have put more patients at risk of death by suicide?
- The authors pose the question: “Should clinicians shift treatment from standard DBT to DBT-S?” How do you answer this?
- What does this study say about how DBT could be effectively applied in treatment settings with limited resources, like public mental health clinics? Does it inform the practice of individual providers in communities with limited access to standard DBT?

Post-Guide

M.M. Linehan *et al*. (2015). Dialectical Behavior Therapy for High Suicide Risk in Individuals with Borderline Personality Disorder: A Randomized Clinical Trial and Component Analysis. *JAMA Psychiatry* 72(5): 475-482**.**

Take Home Summary

Many psychiatrists find borderline personality disorder (BPD) to be difficult and frustrating to treat. Much of the standard practice of mental health care can be subverted by severely borderline patients who will escalate their maladaptive coping strategies in response. To address this, Marsha Linehan developed Dialectical Behavior Therapy (DBT). DBT is based on the concept of *dialectics* (pairs of opposing concepts such as “acceptance and change” that must be brought into balance) applied with an intense structure that involves both group and individual therapy, high access to therapists via phone or email between sessions, and generous peer support for the therapists. Thus, DBT programs are expensive to run and can accommodate relatively few patients at once. This has limited access to DBT, while evidence has built that it is the most effective treatment for BPD and also can be helpful for patients who may not have BPD as their primary diagnosis but also need the skills presented in the program. This study asks whether there can be benefit from isolated aspects of a DBT program so that access to DBT can, in essence, be expanded.

This article describes a randomized clinical trial comparing “standard DBT” to two different interventions composed of elements of DBT, individual therapy and DBT skills group. The primary outcomes were suicide attempts and non-suicide self-injury (NSSI), but the authors also tracked utilization of higher levels of mental health care (e.g., hospitalization), symptoms of depression and anxiety, and suicidal thinking. Study participants were women aged 18-60 with borderline personality disorder, a history of repeated suicide attempt/self-injury, and suicide attempt/self-injury in the last 8 weeks. The authors hypothesized that standard DBT, which contains both elements, would be superior to either of the components alone. Subjects engaged in their assigned treatment for an entire year and were followed for an additional year to examine relapse behaviors. To help balance the amount of exposure to therapists subjects received, the two arms that included just one component of DBT offered ‘placebo’ forms of the other component, individual or group meetings that did not use specific DBT elements but were design to simply provide supportive face time with a therapist.

The primary analysis found no differences between groups on the rates of suicide and self-injury other than a higher frequency of episodes of self-injury among patients in the individual therapy arm. The individual therapy group also demonstrated a higher number of ED visits and psychiatric hospitalizations in the follow-up year after active treatment concluded. Individual therapy also seemed less effective in reducing depression and anxiety during treatment, but those patients had caught up by the end of follow-up. All three arms improved significantly on all of the metrics during the trial (this data is presented in the supplemental information), so it seems reasonable to conclude that although the effect of individual therapy appeared to be less than skills training, DBT and its components are individually effective. This suggests that clinicians should feel good about access to any DBT component, but especially skills groups, for their patients.

Technical Point:

Stratified randomization can be confusing – isn't randomization supposed to even out confounding variables? For those with a little math background, you may recall that 'random' doesn't mean evenly distributed, but instead results in a Poisson distribution, that as the N grow larger and larger approaches a flat distribution. Poisson distributions involve clusters, so sometimes there will be confounding variables that end up unevenly distributed across study arms, especially if studies have smaller samples. In very large trials of more than a couple hundred subjects things are usually fine, but sometimes, as in this trial, recruiting and treating a sample that large would be essentially impossible, so randomization may be stratified. Typically this is done by keeping more than one randomization list. So imagine in a study subjects are randomized by flipping a coin (usually it's a computer of course!). Whoever does this then would write down, subject 1 = group A, subject 2 = group A, subject 3 = group B. In stratified randomization by age, subjects are divided by age, so that say those over 40 are on one list, and those under 40 on another, and each group is randomized and kept in its own list. I recognize that this does not intuitively imply that it avoids uneven randomization, because the human brain struggles with probabilities, so I'll give another example, called block randomization. In block randomization, patients are assigned blocks of, usually 8-10 subjects, set so that each block is half group A and half group B, you can think of the process as being pulling slips from a hat in this case. Once 8 people have been enrolled, you start the 8 'slips' again so that your sample becomes a sample of distributions of size 8. By having half the blocks include subjects over 40, and half subjects under 40 you could also stratify by age, making sure your two arms will have similar age distributions.

Pre-Guide

Liu R-J *et al*. 2012. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biological Psychiatry* 71:996-1005.

Reasons for choosing this article

Here it is: a basic science journal club article. We expect this to be more difficult to read than most journal club articles. Don't worry if you don't quite follow all of the details. Do your best to follow the argument the authors make. Use the **bold text** section headings as a guide. One thing to note about papers reporting on "basic" research is that they nearly always report on multiple experiments. Unlike clinical research where one trial yields hundreds of papers, basic scientists may condense years of work into a publication for a top journal. That's a big part of why basic research papers are difficult to read. It can be helpful to look at the figures to keep it all straight as typically each one will correspond to a single experiment.

- This article was published in the journal *Biological Psychiatry*, which publishes basic science and clinical research and seeks to be read by clinicians. This journal is a good place to get a sense of cutting edge research with strong implications for psychiatric disease.
- This article touches on several concepts that it's good to be familiar with: (1) the idea that BDNF is important for the growth of synapses (synaptogenesis) which is in turn important for depression; (2) investigation of the interaction between medication effects and gene polymorphisms, with the ultimate goal of targeting therapies based on genetics.
- This article also seeks to investigate the mechanism of ketamine, which is being heavily studied for use in patients with treatment resistant depression.

Background

- What human disease do the authors seek to understand?
- Why are the authors interested in this particular BDNF mutation? What is known about carriers of the less common Met allele in humans?
- What is the proposed connection between BDNF and ketamine?
- What was the hypothesis?

Methods

- What types of mice were used in the study? Why were these groups compared?
- What techniques did the authors use? What type of questions did they ask with each technique?
- What is the forced swim test? What is the importance of immobility in the forced swim test?

A technical point from the Methods:

- Genetic models of disease using mice are very common. Why do so many experiments use genetically modified mice? What do “knock-in,” “knock-out” and “wild-type” mean?

Results

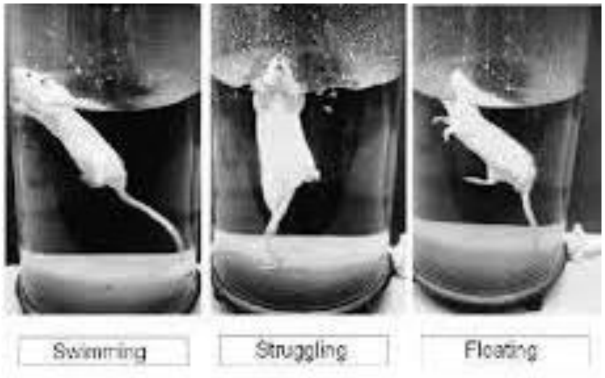
- What were the differences in dendrite structure between the different genotypes?
- What were the differences in response to serotonin and hypocretin?
- What changes in synapses were observed after ketamine exposure in wild-type mice (Val/Val) mice? How was this difference in the other genotypes?
- What changes in behavior in the forced swim test were seen in wild-type mice after treatment with ketamine? In the other genotypes?

Discussion

- What do you take away from this study?
- What is the importance of the observed changes in dendrite structure and response to serotonin and hypocretin observed in the Val66Met mice?
- What is the connection between ketamine, BDNF, synaptogenesis, and antidepressant efficacy, based on the data presented here?
- How would you figure out if the mechanism presented here occurs in humans? Why is it important to understand this?
- What are the (future) clinical implications of this work?

Extra Info for the Perplexed Regarding the Methods:

They use two main methods: 1) in vitro experiments in brain slices and 2) behavioral testing in live mice. In the first set of experiments they take brains from mice with all three genotypes and compare the anatomy of the neurons and their firing without and with neurotransmitters infused into the slice. For these studies, the mice are killed, the brain is rapidly removed, and a specific slice of brain tissue containing the PFC region of interest in placed in petri dish containing a fluid the replicates CSF and has oxygen bubbled into it. This procedure keeps the sliced “alive” for a few hours for the conduct of the studies. The authors then insert electrodes (in a very specific way) into the slice so that they can record electrical currents that reflect the firing of axons (excitatory post-synaptic currents, EPSCs). They do this first at rest and then record changes in the



current after adding serotonin or hypocretin. Recording of the current are shown in Figure 3A. The authors then use very fancy microscope techniques to capture images of the structure of the dendrites of neurons in the PFC area of interest. Once they have the images, they count and measure different areas to quantitate the size and shape of the dendrites (images are shown in the first panels of Figure 1A and 1B, with the graphs showing the quantification). In the live mouse experiments they use the Forced Swim Test (FST) to test antidepressant effects of ketamine treatment. The FST assumes that a "depressed" mouse will give up trying to escape from a container of water sooner than a "healthy" mouse. Mice are placed in a tall, thin beaker in which they can't reach the bottom—thus, the only things they can do are swim around to try and find an escape or float with their noses sticking up out of the water so that they don't drown (rodents are excellent at this!). The amount of time they spend swimming (mobile) is compared to the amount of time they spend floating (immobile) during a 5 minute test. This test is interpreted with immobility equivalent to "giving up," and antidepressants that are clinically effective decrease immobile time in the FST. This test is commonly used because rodents given antidepressants have statistically longer swim times, even though its not clear it really says much about depression. You may hear the FST described as assessing depression in rodents—this is not accurate. The human disease of depression has many different symptoms and effects many different behaviors, while the FST discreetly assesses one behavior.

Post-Guide

Liu R-J *et al*. 2012. Brain-Derived Neurotrophic Factor Val66Met Allele Impairs Basal and Ketamine-Stimulated Synaptogenesis in Prefrontal Cortex. *Biological Psychiatry* 71:996-1005.

Article Summary

The purpose of this article is to ask whether a mutation in the gene for Brain Derived Neurotrophic Factor (BDNF) affects the ability of neurons in the cortex to form synapses. This is clinically important because this naturally occurring human mutation (polymorphism) is common and may be a contributing factor in the development of depression. Patients with this mutation may also respond differently to treatment than patients with the more common allele.

The researchers have previously show that synapses in the medial prefrontal cortex are less well developed and active in mice that have been stressed and that ketamine treatment reverses this. This past work, however, did not define the mechanism by which the synapses malfunction. However, BDNF is already known to be important in maintaining normal synaptic function. This makes BDNF a prime mechanistic suspect – however since mice don't naturally carry the BDNF mutation, it had to be "knocked in" to their genomes for study. In humans, the 66th amino acid in the BDNF protein may be either a valine (Val) or methionine (Met). The presence of Met in this position is less common and makes the protein less active. People who inherit a VAL producing genotype from one parent and the MET producing genotype from the other are VAL/MET heterozygotes and have an intermediate phenotype. The authors hypothesized that the effects of ketamine would be attenuated in mice carrying MET alleles because ketamine is thought to rely on BDNF for its action in reversing depression.

The authors observed significant signs of poor neural and synaptic growth in the mice with MET/MET genotypes and less severe differences in VAL/MET mice. They also found that the activity of neurons when stimulated with serotonin or orexin (also called hypocretin) was decreased with MET alleles, which showed that there were functional differences associated with the anatomical differences. In live animals they found that while ketamine caused increased growth of synapses in VAL/VAL mice, mice with MET alleles had much less growth in response to ketamine, supporting the idea that ketamine relies on BDNF. Similarly while VAL/VAL mice showed significantly less immobility when receiving ketamine in the FST, there was minimal effect in VAL/MET mice and no effect in MET/MET mice. This suggests that the mutation blocks the effect of ketamine as an antidepressant. The authors conclude that their results support the overall model of how BDNF supports normal mood and brain function and how ketamine causes antidepressant effects in a BDNF dependent mechanism.

Comments

Only about 5% of human beings are homozygous for the MET allele of the BDNF gene, but about 25-30% carry a single MET allele. In addition to depression this BDNF mutation is associated with lower IQ and with obesity and related metabolic problems. It's at a scientific sweet spot of being clinically significant and not too rare, making it one of the most heavily studied gene mutations. BDNF has been shown to be a required component of response to antidepressants in general, not just ketamine, but rapid action of ketamine makes it especially suited to explore the role of BDNF in mood regulation. Given the excitement around ketamine as a potential treatment for patients with poor response to traditional antidepressants, this paper also says something important about the use and limitations of this new kind of antidepressant.

Technical Point

The use of genetically modified animals for scientific research is extremely common. Mice make up the vast majority because they have a higher than typical (even eerie) tolerance for being inbred. There are a number of strains sold and each strain is so genetically uniform that the animals are almost clones. The investigators know that when they mutate the BDNF gene, this is the *only* genetic difference between their groups. Combined with the ability to feed and house animals identically, mice show much less variability in experimental responses than humans and other animals (including rats, which are generally outbred). This allows the use of a relatively small number of mice. However, this uniformity may "trick" us when it comes to translating research findings which may not hold true when moved into genetically diverse populations. There are several methods used to insert, remove or change DNA so that the mutations are passed down in the offspring. In general, if DNA is added to an animal, it's called a "knock-in" and if its deleted a "knock-out." In this case, there is only a sequence change but the investigators choose to use the term "knock-in" to describe the animals. Animals that have not been manipulated are called "wild-type" even though they are far removed genetically from the mice in true wild populations.

If you were wondering whether the finding that MET carriers are less likely to respond to ketamine has been replicated in humans, the answer is… maybe? A small study in humans showed some difference in effect by genotype but no large and definitive studies have been published.

Pre-Guide

EE McGinty et al (2016). Trends in News Media Coverage of Mental Illness in the United States: 1995-2014. *Health Affairs* 35(6):1121-1129.

Reasons for choosing this article

- The topic of the article—the portrayal of mental illness in culture—is worth examining.
- The article gives an example of healthy policy research, which is an important type of research that we don't have as much opportunity to think about.

Background

- The authors argue that how mental illness is portrayed by the news media is important and has real world consequences. Why do they think this is important? Do you agree?
- In your own life, what do you notice about media reports that include mental illness? Have you had conversations with family or friends (non-psychiatrists) about topics like mental health care policy and the role of mental illness in violence? Have media reports informed those conversations?
- What knowledge gap is the study designed to fill?
- What (if any) hypothesis do the authors have for the study?

Methods

- What data were studied? How did the authors choose data for inclusion?
- How many articles were analyzed? Does this seem like an appropriate number to you? Why or why not?
- What was the proportion of print to television stories analyzed? Does this seem appropriate to you?
- What measure was used to categorize the news story content? Was the measure validated in any way?

A technical point from the methods:

- In the measures sections of the methods (page 1123), the authors describe developing a "sixty-nine item structured coding instrument" and later in the same paragraph refer to "six specific consequences of mental illness were coded." In these sentences, what do the words "coding" and "coded" refer to? What does this mean that the authors did in practice? What steps did they take (or, are typically taken) to ensure the reliability of the coding?

Results

- What topics were most commonly discussed in media reports about mental illness (Figures 1 and 2)? Were there changes over time? Does anything surprise you about the commonly discussed topics?
- Why do you think the authors included the dates of several mass shootings in Figure 1? Do these events appear to affect media reports?
- Per media reports, what are the common causes and consequences of mental illness (Figure 4)?
- How much coverage of mental health policy (Figure 5) did the authors observe? How did this compare to the number of stories about violence (Figure 2)?

Discussion

- What do you take away from this study?
- How does the media portrayal of people with mental illness match your understanding of people with mental illness? Do you think the media is accurate in its portrayal of mental illness?
- What are the limitations to this study? (The authors provide their list of limitations at the end of the Methods, which you can agree/disagree with or add to.)
- The authors state "these findings raise troubling implications for social stigma toward people with mental illness." Agree or disagree?
- What is the role of physicians in media reports on mental illness? Should psychiatrists take an active role in discussing mental illness in public spaces? Advocate in other ways?

Post-Guide

EE McGinty et al (2016). Trends in News Media Coverage of Mental Illness in the United States: 1995-2014. *Health Affairs* 35(6):1121-1129.

Summary

It may often seem to psychiatrists that our field suffers in the public imagination. Stigma affects our patients directly but also be self-reinforcing through cultural and political neglect. In order to better understand the view of mental illness in society, it is important to go deeper than our own, possibly biased, perceptions and gather data. It is in this spirit that the authors of this paper performed an analysis of media depictions of mental illness.

To do so they performed a review of media from 1995-2014, which consisted of print and some television coverage in the form of transcripts. The databases used to obtain media stories contained national and large-regional media sources but not local papers or television news. Articles were searched by headline and then reviewed by the authors to classify their content. The primary goal was to assess the rates of various types of content in media, which the authors divided into five area: topics (such as suicide), consequences (such as being arrested), causes (such as stress), individual depictions (such as a story about a patient entering recovery), and policies (such as legally mandated substance abuse treatment). Each of these was further subdivided into specific categories. Each story could be assigned to as many bins as applied to its content. In addition to determining rates of content, they also looked at trends over time by comparing the first decade (1995-2004) to the second (2005-2014). Of note, the authors did perform statistical analysis for significance on the data set, although the focus is on descriptive statistics.

The results show that of the topics identified, violence was the most common, followed by treatment and insurance coverage. A minority of articles addressed the science of the brain and behavior. Violence included both self (suicide) and other directed, with violence towards others being the most common topic. A further examination of this result revealed it was related to coverage of shootings, especially mass shootings, with spikes of articles being published after a shooting that mention possible or presumed mental health problems in the perpetrator. This pattern, with stories about violence mentioning mental illness, is seen throughout the entire analysis. The authors found that while 38% of news stories about violence that mentioned mental illness stated that mental illness increases the risk of violence, only 8% mentioned that the (vast) majority of people with mental illness never commit a violent act. Similarly, few articles mentioned other possible causes of violence.

A minority of stories mentioned any basis for mental illness, with stress being the most common, closely followed by biology. About half of the stories discussed a consequence of mental illness with incarceration or other criminal justice system involvement being the most common. Although almost half of the stories included specific stories about an individual, four times as many described violent acts than successful treatment. Finally, mental health policy was rarely brought up, but when it was, much of it called for improved care and access. Generally, the findings were stable over the two decades analyzed, though there was a trend to increasing coverage of causes of mental illness, and interestingly, less coverage overall.

The results overall confirm the impression that media coverage of mental illness is focused on inaccurate violent portrayals of the mentally ill, but they also suggest that this may be because acts of violence are the most likely reason that mental illness is mentioned in the media. This raises the possibility that increasing positive depictions of mental illness – or even any coverage at all – between episodes of violence could make an impact on public views of the mentally ill and illness origin, prognosis, and treatment. It should be noted that the sample was biased towards media with wide coverage, so that if local media have a different distribution of content, the overall exposure captured may not be representative. Additionally, the list of topics was chosen by the investigators and its unclear whether some important areas of mental health coverage may have been left out of the analysis.

Technical Point

Data coding: In order to perform statistics on ideas those ideas must, in some way, be made into numbers. To do that investigators develop standardized ways of collecting and organizing data. You are already familiar with commonly used scales such as the PHQ and QIDS but sometimes no instrument exists or can feasibly exist for a particular project. Assessments like this, with large, atypical data, and fairly specific and unique research questions, are good examples. In this case the investigators used an accepted procedure for data collection and coding. Coding means to assign a category code or codes (historically a number though many kinds of statistical software can now use text codes) to each article, patient, event, etc.

To code the media data, the investigators first developed the list of topics, causes etc they were interested in and gave each of these an official definition shared within the group. For example, they would agree how to classify suicide threats vs attempts vs completed suicide. They would also try to establish benchmarks for when to consider a story, for example, to provide a diagnosis for its subject. Typically, once the definitions are finalized, and a list of material is selected, the coding is done by at least two people for each item (in this case, each article). Each person builds a spreadsheet of codes for each article and then these are compared with the codes given by the other coder. Any differences are resolved by a pre-determined process. Then a ‘master’ data set it created with the final codes.

All of this is a lot of work! If you are imagining days or weeks of reading and coding, that’s probably accurate for an article like this. However, it adds a lot of credibility to the analysis to go through this process, and any good paper converting soft data into numbers should give you a window into how this was done in the methods. You should also understand how “human” this whole process is, and recognize the need to see details of the process to judge the quality of an analysis.

Pre-Guide

DE McNiel and RL Binder (2007). Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *American Journal of Psychiatry* 164: 1395-1403.

Reasons for choosing this article

- This article offers an introduction to the idea of mental health courts, an important aspect of community psychiatry.
- This paper seeks to determine the effectiveness of an intervention without using a randomized controlled design, which allows us to consider times when an RCT is not the appropriate method for answering a question.

Background

- What sorts of things are the authors referring to when they note the “large-scale involvement of people with mental disorders in the criminal justice system”? In what settings have you seen patients with serious mental illness who were involved with the criminal justice system? What diagnoses were typically seen among these patients? What types of criminal charges?
- The authors note that mental health courts have therapeutic goals. Do you think this is an appropriate use of the criminal justice system?
- What hypotheses do the authors have for the study?

Methods

- How did the authors identify the study participants?
- The authors note that “informed consent was not required.” Why not?
- How did the authors match the treatment as usual and the mental court participants? Why did they include so many treatment as usual participants?
- What were the study outcome measures?

A technical point from the methods:

- In your own words, describe “propensity scoring.”

Results

- How did the mental health participants and treatment as usual participants compare on the baseline variables assessed in Table 1? What does it mean that some values in Table 1 have a significant *p* value?

- What was the effect of mental health court participation on crime recidivism while in the mental health court program? After the program was completed?

Discussion

- What do you take away from this study?
- Pretend the district attorney's office has asked you if Dallas should institute a mental health court program. What advice would you give? What parameters would you give for which people should be offered the chance to participant in the program? Would the type of charge or diagnosis be important considerations?
- In the background, the authors provide their rationale for why the used the design they did and why a randomized controlled trial would not be appropriate way to determine the effectiveness of mental health courts. What do you make of these arguments? Do you think the study provides an answer to the authors' question?

Post-Guide

DE McNiel and RL Binder (2007). Effectiveness of a Mental Health Court in Reducing Criminal Recidivism and Violence. *American Journal of Psychiatry* 164: 1395-1403.

Summary

This paper addresses an important issue in mental health that many providers don't often think about unless they specialize in it: forensics. There is a long and often shameful interaction between the legal system and the mentally ill. Historically jails and prisons have had a significant role as a place for the mentally ill to reside in society, and in our current system and cultural moment, they do again. Unlike in past times, however, there is a recognition of the high rates of mental illness in the prison system and effort towards addressing the needs to these offenders for the betterment of both their lives and society.

Specifically this paper addresses the impact of parallel legal infrastructure for the mentally ill – courts with staff who specialize in the needs of offenders who are also patients. The authors examine data from the mental health court in San Francisco from 2003 to 2005. The explicit goals of the program were to increase treatment for the mentally ill offenders by considered access to care in sentencing, with a goal of decreasing rates of return to the legal system. Therefore the authors assess whether this goal was met by the program.

In order to participate in the mental health court system, offenders were diagnosed with a "severe" mental illness and provided consent for treatment in the community mental health system. Although the definition of severe mental illness is not made explicit, but information about specific diagnosis is available in the footnote to table 1, showing that psychotic disorders accounted for most of the mental health court group, while only a small percentage of those in the treatment as usual group (<20%) had psychosis. In contrast rates of substance use disorders were fairly consistent between the groups (~60%). Several other variables were substantially different between groups (assignment to the mental health court system was intended to be non-random); the technical point will discuss how these differences were handled by the authors. A baseline period of 12 months before enrollment into the mental health court (or prior to the first arrest during the data period) was compared for each subject to the 12 months after this index event. The primary outcome was rate of re-arrest, split by violent and non-violent crimes.

Using a survival analysis, the authors found that time to re-arrest was longer while enrolled in mental health court and for up to 24 months after "graduation." The size of the effect was about 20% at 24 months, which is a meaningful difference. The authors concluded that the court was effective at its stated goal. Although the efficacy of programs like this is often measured in terms of money – a subject not addressed here – in terms of personal and public safety and health, mental health courts represent one of our better approaches.

Technical Point

Propensity scoring is a way of attempting to correct for non-random assignment that affects observational studies. Essentially they first use the data to come up with the probability that any given person would be assigned to one group (in this case the mental health court group). Then the weight that subject's data is given in the analysis is adjusted based on the probability; in general those with very high probability of assignment to one group are weighted less than those whose probability is closer to 50:50. In this case they essentially performed an entire regression analysis to identify which variables predicted court assignment – they were: race, homelessness, diagnosis, early entry into the program (when the court was new, presumably it enrolled at a more rapid rate), total charges in the last year, and violent charges in the last year. Once these variables were identified, they were able to use the coefficients to calculate the probabilities, convert those to weights, and then compute a weighted Cox model of the results. This method is capable of modelling results from studies in which participants are randomized but the success at doing so requires the sample be sufficiently large, and that two groups have some characteristics in common – that is if 100% of the subjects with severe mental illness were sent to mental health court, it would not be possible to compensate statistically. Additionally, like all models, variables which were not measured cannot be accounted for, and can lead to misleading results.

Pre-Guide

IW Miller *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Reasons for choosing this article

- This article addresses a common clinical scenario: patients seen in the ED who report recent history of suicidal ideation or attempt.
- The trial is not a randomized controlled trial, which lets us think about how to questions that may not be ethically appropriate for an RCT.

Background

- What data suggest that universal screening for suicide is a good thing to do? Do you agree with the authors' statement that emergency departments are "particularly important locations for suicide prevention?"
- What typically happens when a patient screens positive for suicide in the medical ED?
- What was the study hypothesis?

Methods

- How were participants identified? Do you consider the inclusion and exclusion criteria to be appropriate, given the study hypothesis?
- What happened in each phase of the study?
- What were the components of the intervention? Do you think this is scalable/generalizable?
- What was the primary outcome? What tool was used to assess this? (hint: https://cssrs.columbia.edu/wp-content/uploads/C-SSRS-1-14-09-SinceLastVisit_AU5.1_eng-USori-1.pdf)
- The authors given their power analysis, and they state that the studied was powered to detect a 7% absolute risk reduction in the rate of suicide attempts. Is a 7% absolute risk reduction clinically meaningful? Would you consider a decrease smaller than 7% to be meaningful?

A technical point:

- How would you describe the type of study the authors conducted? Why might the authors have chosen this design and not a randomized clinical trial?

Results

- Review Table 1. How do the study participants compare to your patients (or to the patients who've seen on consults who screened positive on the universal suicide screening)?
- How often was the study intervention actually done? What do you make of this?
- Did screening affect the rate of suicide attempts? Did the intervention affect the rate of suicide attempts? (Figure 2)
- The authors report a number needed to treat (NNT) of 22 for the intervention. What does this mean? Is 22 good, bad, somewhere in-between?
- Are you concerned that there were more numerically more completed suicides in the intervention phase than in the treatment as usual phase? Why or why not?

Discussion

- What do you take away from this study?
- What are the barriers to implementing the intervention from the study? Given these barriers, should we be doing universal screening for suicide in medical EDs?

Post-Guide

IW Miller *et al*. Suicide Prevention in an Emergency Department Population: The ED-SAFE Study. *JAMA Psychiatry* 2017; 75(6): 563-570.

Take Home Summary

Suicide is one of the most intractable problems in Psychiatry, and like many 'rare' events is difficult to study in clinical trials. As far as we have been able to tell, little physicians and the health system do changes the rate of death by suicide. The study reported in this article takes a non-traditional approach to suicide reduction. Although many patients who make a suicide attempt have contact with a health-care provider of some kind in the days before the attempt for many of these patients the contact is not with a psychiatrist or therapist, it's with non-specialized emergency services. So the ED-SAFE study was designed to utilize that point of contact for suicide reduction.

The ED-SAFE study was designed around two questions – does active ER screening for recent suicidal ideation and behavior impact rates of such behavior *after discharge*? And can interventions reduce such behavior? The study was a quality improvement project more than a traditional piece of clinical research. The investigators assessed baseline suicide attempt rates, and designed and implemented two phases of system change in the 8, non-specialized ERs that participated in the trial, and then tracked the results of those operational changes.

In the first phase of the study the investigators just tracked data on patients who presented to the ER and voluntarily told the teams caring for them that they were suicidal or had made an attempt. While some clinicians at some sites might have asked about SI in the patients, others certainly didn't; this established the treatment-as-usual baseline. In the second phase, all the sites implemented universal screening. This meant that in theory some patients 'missed' in phase I were picked up in phase II, and indeed the rate of patients identified as meeting criteria for enrollment increased in phase II (detail of this are the focus of a prior publication). In the third phase, an intervention in which subjects received a more detailed assessment from a physician, a crisis planning session, and periodic contact by phone. The intervention focused on helping patients identify resources and create plans rather than on typical therapeutic interventions as would be found in CBT or DBT. Outcome data was also collected by phone and while suicide attempts were the main outcome they also looked at a composite outcome that also included other suicidal behaviors (but not ideation alone).

The study found that about 20% of patients enrolled made a suicide attempt in the year after enrollment. Screening alone did not significantly change the rate (22.9% vs 21.5%) but the intervention did (18.3%). The results were similar for the composite outcome. The intervention appears to be well liked, of those who ever engaged in a phone call, most stayed in contact for several additional calls. The relative reduction in risk was 20% for attempts and 13% for the composite measure, both a clinically meaningful amount. Before the study began few would have argued against the practice of increasing screening and intervention for suicide, but given the operational costs of such programs, it was necessary to have strong evidence that they make a difference on a systems level. The study successfully showed than a relatively low-cost and easy to implement intervention can make a meaningful change in the long term behavior of patients.

Technical point

Over the past few decades, allopathic medicine has made a lot of progress towards truly embodying the principles of evidence based medicine. As a result, there has been something of a fetishization of the Randomized Clinical Trial (RCT, capitalization intended). It may sometimes seem (and has been argued) that without an RCT interventions shouldn't be recommended in guidelines and public health recommendations. Recall, however, that RCTs make up one of many forms of evidence, and while they are the best way of addressing many clinical questions, there are times in which they are impractical, unethical, or impossible (don't forget about, for example, the *entire field* of epidemiology).

In clinical intervention research, often the reason why an RCT *shouldn't* be done relates to a concept called equipoise. Equipoise means that, at the level of expert consensus of the available evidence for an intervention (or in the case of a trial, the pairing of the trial arms) *no recommendation can be made*. In the abstract, this can be reduced to 'don't experiment on human beings if you already know the answer' but, of course, usually it's complicated to determine in practice. Often some experts recommend an intervention, while some recommend against it; if these are approximately equally split, then equipoise exists. If only a few people are on one side, then it does not. Sometimes rather subtle distinctions between when and how exactly an intervention should be delivered and obscure the situation entirely until more work is done. Understanding equipoise significantly enlightens, for example, the use of placebos, by reminding us that we must lack evidence that a drug will have more benefit than harm to use a placebo in a trial.

In this case, we have trouble thinking of a way in which the screening or interventions proposed in the ED-SAFE study could harm. In this sense, a traditional RTC in which patients were randomized to treatment as usual, screening, or screening + intervention would not be in equipoise; we were already pretty sure that we could do better than treatment as usual. The investigators also considered that the interventions were operational changes very difficult and costly to implement in large naturalistic clinical settings. It was important, for example, that the intervention be provided by the doctors, nurses and staff in the actual sites rather than by trained research personnel; showing that the sites could do the intervention was a significant part of the trial. In many ways ED-SAFE was more a big quality improvement project than a traditional study. This suggests that one option could have been to randomize *sites* rather than patients. One design often employed in studies with operational interventions is the "step-wedge" design, in which all sites start collecting treatment-as-usual data, and then start the intervention in randomized order. That way, all sites receive the intervention but there is ample collection of control-condition data. However if it's feasible for all the sites to start the intervention at the same time, then a step-wedge design may not offer benefits over the sequential design used in ED-SAFE.

Pre-Guide

Olson K.R. *et al*. 2016. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics*. 137(3): e20153223.

Reasons for choosing this article

- This article provides information on a topic that is of great interest in the popular press: what are the mental health effects of social transition for transgender children?
- This article lets us talk about "negative data"—studies that do not find differences between groups.

Background

- Have you had any clinical experience with transgender patients? In what settings?
- Why has previous literature focused on transgender adolescents and adults but not children?
- Why do the authors include Table 1?
- What would have been a reasonable hypothesis for this study?

Methods

- Who were the study participants? How did the authors find them?
- Who were the controls? Why do you think the authors included 2 control groups? What are the advantages and disadvantages of using cisgender siblings as control?
- What scale did the authors use to measure symptoms? What do you think of the use of this questionnaire? What are the advantages/disadvantages of using symptom checklist vs use a tool or interview for formal diagnosis?
- There are times when the authors use the term "depressive symptoms" and times when they use "depression." What is the difference between the terms? Given the measures they used, which more accurately reflects what the study measured?

A technical point from the Methods:

- The authors report the following on the scale they used: "scores are nationally normed and provide a t-score such that score of 50 represents the national mean, with a SD of 10." What does this mean?

Results

- What comparisons did the authors make?

- The authors report a higher level of anxiety symptoms in transgender children compared to the national norm, but they also note that this level of symptom is below the preclinical level. What do you make of this finding? Do you think it's important?
- In Table 3, all of the *p* values are greater than 0.05. What does this mean?
- What are the major findings of the study?

Discussion

- What do you take away from this study? What is the importance of these results?
- Typically, authors find it very difficult for high quality journals to publish studies with "negative data"—studies that don't find differences between groups. What do you think makes journals resistant to publishing negative data? Why do you think the data presented here were accepted for publication?
- The authors note that "one might reasonably ask whether this study provides support for all children with gender dysphoria to socially transition." What do you think? What factors may limit the generalizability of the findings?
- In discussing the limitations of the current work, the authors note that this study is not a randomized controlled trial. Would it be possible to answer the questions posed by this study with an RCT? What would that design look like?

Post-Guide

Olson K.R. *et al*. 2016. Mental Health of Transgender Children Who Are Supported in Their Identities. *Pediatrics*. 137(3): e20153223.

Article Summary

This study attempts to fill in the large gap in what is known about outcomes for children who are transgender. Although overall the methods, sample, etc., are weaker than in most articles we chose for journal club, the lack of validated information in this area increases the study's importance. An important related feature of the article is the careful way in which definitions are laid out, despite the fact that the intended audience is clinicians, who in most cases would not need this information. For example, it's important to understand that being transgender is not the same thing as having gender dysphoria.

This is, on the surface, a simple comparison between transgender children and controls but it is more complicated upon further examination. It is debated whether to compare these children – clearly far from a random sample – to the American population as a whole (at least as much as possible) or to children their similar cisgender siblings. Furthermore, while the authors take great care not to explicitly predict that socially transitioned children will have normal mental health, this prediction/hypothesis remains unspoken and significantly affects the analysis. If that is their prediction, then an equivalence analysis would be appropriate (I feel that this is the direction the authors should have taken but it is legitimately debatable). Instead they cite literature showing high rates of mental illness among transgender youth and use a standard analysis or difference. An additional feature of significant interest is the use of normalized symptom assessments. These provide data that are highly processed to reflect the US population – in this case considering age and other factors. This is probably useful because it provides a better "general" control comparison, if that is desired, and is notable, because despite the government resources providing such norms, this method is used relatively rarely.

It's worth spending a few moments looking at the demographics table, given that the sample was self selected (that is, families that chose to socially transition their children). The sample is primarily white and overwhelmingly of higher socioeconomic status. We imagined these families as also having higher education (though this data is not reported) and probably taking an assertive advocacy stance for their children. It's also interesting to note the gender skewing, with far more boys-transitioned-to-girls than the other way around. We do not know if this is representative of the transgender community as a whole or was specific to this study; they authors make no comment on it. Comparing the scores on the mental health measures, the authors found that transgender children have scores close to US norms on both measures. There was also no significant difference between transgender children and either group of control children, although the difference in anxiety was close to significance. The authors then present a table comparing scores in this sample to two prior studies but without performing statistical comparison. This is very, very unusual. In this case, it probably only passed peer review because there is not enough existing data to perform a statistical analysis (i.e., a meta-analysis).

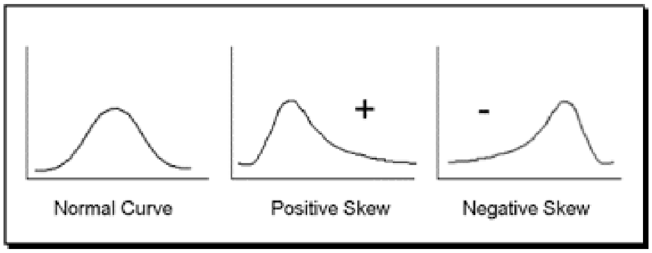
Comments

This is a rapidly evolving field of study and clinical area. It also touches on many social and political controversies in the US right now. As clinicians it's important for us to advocate for the well being of all our patients. We think this study sets an important precedent in terms of support for making gender transition early in life for those kids who can be identified as truly transgender. At the same time, there are serious concerns about whether or not the sample can be considered representative. The children featured here likely have social support and more resources than typical kids, and may simply be more resilient, and have families that are more resilient, to the stresses of being transgender. Nonetheless, we can expect more studies like this to be done and used to support policy regarding sexual and gender differences in the US population.

Technical Point

In this study the authors used normed scales. The most well known example of a normed scale is that for IQ. When you take an IQ test, scores are reported so that 100 means a mean score. Most people don't realize it, but your IQ has probably changed over your lifetime because IQ scores are renormalized to the population, and the mean performance has steadily improved over the last few decades. Although the measurement of IQ has been controversial, there are benefits to using scales with population norms. First they help you know that your measurements in your sample are reliable. It may happen if you have staff that are particularly adept at extracting symptom reports, for example, you might get inflated scores. Poor instructions to subjects may also result in scores that are consistently too high or too low.

Another reason these scales are helpful is that they allow us to compare studies to each other more easily, because we have an idea of what normal is. A limitation is that you can't really assess subjects the way we do as physicians. There is an inherent conflict between detecting an illness condition such as depression and measuring symptoms, which in isolation may not be specific or clinically relevant; scores may correlate to a diagnosis but will never perfectly align. This is reflected in the score cut-offs given as "clinical" range scores. On a related note it's difficult to understand whether differences in "depression symptoms" far below the "clinical range" have any meaning at all. That is while the difference between 60 and 70 on the scale may be several more positive answers, a difference between 40 and 50 is likely to be, essentially, zero. It depends on the nature of the underlying curve. Qualities such as IQ which tend to fall in a normal distribution are better suited to normalized scales than those like depression symptoms, which tend to be right (positive) skewed, see figure at left.



You can find out more about the PROMIS measures at: http://www.healthmeasures.net/explore-measurement-systems/promis

Pre-Guide

V Popova *et al*. (2019) Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with a Newly Initiated Oral Antidepressant in Treatment Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *AJP* 176(6): 428-438.

Reasons for choosing this article

- This is a “pivotal” phase 3 trial of intranasal esketamine, which was approved by the FDA for the treatment of depression on March 5, 2019. The approval was covered widely in the popular press, and you may already have encountered patients interested in this new treatment.
- Many of the study authors are employed by Janssen, the subsidiary of Johnson and Johnson that developed intranasal esketamine. This article lets us think about the possibility of conflicts of interest in pharmaceutical development.

Background

- How often do patients respond to the first antidepressant medication with which they are treated? The second medication? Third medication?
- Prior to the approval of esketamine, what medications had evidence for benefit in patients with “treatment resistant” depression? How well do those options work?
- Prior to reading this article, what was your impression of the efficacy of ketamine/esketamine for depression?
- In the background, the authors note that esketamine was “recently approved” by the FDA. If you look at page 437, you will see that this manuscript was first submitted to *AJP* on 2/15/19 and accepted for publication on 3/28/19. FDA approval was awarded on 3/5/19. What do you make of this sequence of events?
- What was the study hypothesis? (The authors do not explicitly state their hypothesis.)

Methods

- The authors describe this study as “active-controlled.” What do they mean by this?
  - Note: the authors of the journal club pre- and post-guides had an extended discussion about the answer to this question.
- Why did the study start with a 4 week screening and “prospective observation” phase (i.e., what was being prospectively observed during this time)?
- Who were the study participants? How did the authors determine that the study participants had “treatment resistant depression?”
- Do you consider the eligibility criteria to be narrow or broad? Are the eligibility criteria appropriate for the goals of the study?

- What group of people determined the treatment response used in the primary outcome? How much interaction did these "blind raters" have with the participants and other study staff?
- At the beginning of the statistical analysis section, the author describe which data were included in analysis. Was their approach consistent with an intent-to-treat analysis? Is this important?

A technical point:

- The authors note throughout the manuscript that this study is a phase 3 trial and compare their results to prior phase 2 trials. What are the roles of phase 1, phase 2, and phase 3 trials in the FDA approval process? What does each type of study seek to establish?

Results

- How many patients were screened and randomized? How many dropped out of the study? What were the major reasons that screened participants were not randomized? Major reasons for drop out? (Hint: there is something missing from this manuscript.)
- How do the number of participants randomized and retained in the study compare to the power analysis?
- Review Figure 1. What effect did esketamine +new antidepressant have on depressive symptoms as measured by the MADRS? Over what period of time? What about placebo+new antidepressant?
- How did the number needed to treat (NNT) for response vs remission compare? How do you interpret these data?
- How do you interpret Figure 2? Which groups of participants benefited from treatment with intranasal esketamine?
- Was esketamine well tolerated? Are any of the side effects concerning?

Discussion

- What do you take away from this study?
- Would you recommend esketamine treatment to a patient? If so, which patients?
- How does esketamine compare to other treatments available for treatment resistant depression?
- Ten of the authors are employees of Janssen Research and Development and hold company equity. Does this affect your interpretation of the data presented? If yes, in what ways?

Post-Guide

V Popova *et al*. 2019. Efficacy and Safety of Flexibly Dosed Esketamine Nasal Spray Combined with Newly Initiated Oral Antidepressant in Treatment-Resistant Depression: A Randomized Double-Blind Active-Controlled Study. *American Journal of Psychiatry* 176(6):428-439.

Take Home Summary

This article was chosen because of the significance of the FDA approval of esketamine for treatment-resistant depression (TRD). Unlike most of the papers covered in JC, this one reports a study performed by a company in the service of achieving FDA approval for a drug, and the majority of the authors are employees of Janssen subsidiaries around the globe. It's unusual for companies to publish in peer-reviewed journals, but the arrival of not just a new drug for depression but a new class may have motivated publication in this case. However there are numerous ways in which this paper may deviate from non-industry funded work, which are worthy of discussion.

Treatment resistant depression (TRD) is an important area of research in psychiatry right now. Depression is the most common single mental disorder and a significant percentage of patients, about 30%, are treatment resistant, leaving a huge number of patients – about 3% of the population – suffering from TRD. Although there isn't an official definition (i.e., in the DSM) most experts consider a patient who has failed two 'adequate' trials of antidepressants to have TRD.

Esketamine is the S enantiomer of ketamine. Janssen has shown that S-ketamine has tighter binding to the NMDA receptor than the R enantiomer, though there is skepticism that this results in greater clinical efficacy (though using one enantiomer does allow the drug to be patented, for more see https://en.wikipedia.org/wiki/Enantiopure_drug). Its mechanism appears to be modulating glutamate neurotransmission, though this too is controversial. The unique clinical feature of ketamine is that it has rapid (within about 4 hours) onset of antidepressant action; additionally esketamine comes in a nasal spray format that allows parenteral dosing. Ketamine has heavy first pass metabolism limiting its potential as an oral medication.

The study enrolled adults with chronic or recurrent depression who were non-responders to their current medication and who were observed for 4 weeks to ensure they met criteria for TRD (medical records and self-report were used to document the prior treatment failure). They were started on a new oral antidepressant and randomized to esketamine or placebo, dosed twice a week for 4 weeks. Subjects self-administered the spray while being monitored in the office by staff. Results showed that the esketamine group had about 4 points more improvement on the MADRS, a common scale for drug efficacy studies, at all time points over the 4 week trial. This effect is smaller than was hypothesized when the trial was designed (according to power analysis in the methods and a brief statement in the discussion), but still significant. The authors attribute this to higher than expected placebo response. There were significant immediate side effects of esketamine on blood pressure and some behavioral effects, but these were short lived and not serious.

There are a few odd things about this paper, especially given the high impact journal in which it was published. Though the title of the study states "active-controlled" most investigators would consider the placebo used here to be masked – that is made to resemble the drug through the use of bittering agent. Most studies of ketamine and ketamine-like drugs using an active placebo used midazolam, a short acting benzodiazepine; probably most scientists would infer this meaning from the title. The authors of this guide discussed whether the term here means the placebo group was also taking a new oral antidepressant, but this is not how "active-control" would usually be used. Additionally, the manuscript leaves out several types of data – a CONSORT diagram showing the study sample over time (drop-outs), how many patients received the higher dose, and any mention of outcomes over the entire 24 week study period. To us these omissions seem atypical, and invite concern that a different publication standard may have applied to this paper than to others.

Technical Point

The process for approval of new drugs in the US is long and complex. Once a compound has accumulated promising data (usually preclinical data) a company or investigator submits an Investigational New Drug (IND) application to the FDA asking for permission for research use in human beings. Three types of clinical trials must be done to submit a complete approval packet. First, in phase 1 drug is given to a small sample of healthy volunteers, in varying doses, looking for toxicity. Though the number of subjects in phase 1 is small they are intensely monitored to collect data about pharmacokinetics and if applicable, pharmacodynamics. If the drug seems safe, then phase 2 studies begin. These are conducted in the patient population of interest, but are designed to assess safety as the primary endpoint rather than efficacy. Often the drug dosing schedule is refined during phase 2 based on the pharmacokinetic data collected in phase 1. Although phase 2 is primarily focused on safety, companies collect efficacy data to decide whether to continue to phase 3.

Phase 3 trials are the main outcome trials for new drug approvals from the perspective of efficacy; while phase 1 and 2 may not use placebo control, the FDA requires this for phase 3. In fact, two phase 3 trials must be 'positive' with the drug showing superior efficacy to placebo for approval. There are also requirements for how efficacy is measured – the MADRS scale, for example is one of a few allowed for depression studies – how the placebo is designed and administered, how the study is powered, and other features of the design. Phase 3 studies are larger, longer, and far more expensive to conduct than phase 1 or 2. Typically the studies are designed to maximize the likelihood of the drug separating from placebo through picking a healthy, relatively 'pure' sample that excludes people with other illnesses and taking a wide variety of medications. Phase 3 trials also often are shorter than real world duration of drug treatment, and though they assess side effects, are not required to track outcomes beyond the duration of the acute dosing phase, unless the FDA specifically requests that this is done. For example, with drugs like esketamine that may have abuse potential, the FDA may require follow up to track reports of substance use disorders. These long term or real world outcomes are provided in phase 4 trials – which can, and usually do, occur *after* a drug is approved.

Pre-Guide

Psychiatric Genomics Consortium. Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature* Genetics 2019; 51:793-803.

Reasons for choosing this article

- This is not an easy paper to read for those who are not immersed in the genetics literature. Do your best to separate the forest from the trees and follow the main points of the article.
    - **We've included references to a guide for the perplexed that we encourage you to read along with this article: TA Pearson and TA Manolio. How to Interpret a Genome-Wide Association Study. JAMA 2008; 299 (11): 1335-1344.**
- This article is a very high quality example of the work being done to better understand the genetics of major psychiatric disorders and picks up on themes in other articles like B-SNIP.
- Articles like this is that these studies tend to get coverage in the popular press, and understanding the method will help you talk about the results with patients.

Background

- Prior to this paper, what did the field understand about the genetics of bipolar disorder?
- Unpack the following statement: "Although modern diagnostic systems retain the Kraepelinian dichotomy between bipolar disorder and schizophrenia, the distinction between the two disorders is not always clear-cut and patients who display clinical features of both disorders may receive a diagnosis of schizoaffective disorder-bipolar type."
- What was the study hypothesis?

Methods (note: the Methods are scattered a bit, with some information in the Results section and some in the Methods section at the end of the paper)

- Who were the study participants? How was diagnosis determined? According to Pearson and Manolio, what sample size is typically used?
- What is the follow up of suggestive loci in additional samples analysis? How is it different from the first analysis? Does this match with the methods recommendations made by Pearson and Manolio?
- What is an "in silico" analysis, and why is it performed?

A technical point:

- What is a genome-wide association study (GWAS)? What are these studies designed to assess, and what are the important controls?

Results

- The first section of the results talks about tests of linkage disequilibrium and a genomic inflation factor. What are the authors trying to measure with these analyses? (These terms are defined by Pearson and Manolio in the box on page 1337 and discussed at the bottom of 1340-top of 1341.)
- What information is presented in Table 1? What does each column and row mean, and what does the bold type indicate? (Meaning of the p-values is discussed by Pearson and Manolio on page 1340 in the 2nd to last paragraph on the page)
- What types of genes were identified by the analysis? What pathways are the protein products of these associated with?
    - What are the steps the authors take to identify the pathways? (This is related to the question above about the “in silico” analysis.)
- In the abstract, the authors state “Bipolar I disorder is strongly genetically correlated with schizophrenia, driven by psychosis, whereas bipolar II disorder is more strongly correlated with major depressive disorder.” Do the data presented in Figure 2 support that statement?
- Explain the following statement from page 796: “We note that significance levels were assigned to genes by the physical proximity of SNPs, and we do not imply that significant genes are causal for BD.” (Discussed by Pearson and Manolio in the final paragraph of page 1342.)

Discussion

- What do you take away from this study?
- Do you agree with the assessment of Pearson and Manolio that “these studies are clearly many steps removed from actual clinical practice?”
- What are the reasons to do this kind of work? What is the future promise of this study (and others of its kind)?
- The final sentence of the paper is that the results “reveal an extensive polygenic genetic architecture of the disease, implicate brain calcium channels and neurotransmitter function in BD etiology, and confirm that BD is part of a spectrum of highly correlated psychiatric and mood disorders.” What data in the study support that statement?

Post-Guide

Psychiatric Genomics Consortium. Genome-wide association study identifies 30 loci associated with Bipolar Disorder. *Nature Genetics* 2019; 51:793-804.

Take Home Summary

This is a recent, 'state of the art,' genomics paper. It is a meta-analysis of a Genome Wide Association Study (GWAS) comparing people with bipolar disorder to those without. A GWAS looks at gene variants across all chromosomes to identify those that may be causal for a disease. In the early days of GWAS there was a lot of success identifying genes responsible for classic Mendelian diseases. Since then, however, the method has been applied to common disorders with complex heritability (like mental illnesses) with considerably less success. As a result, the methods have gotten more rarified and the samples much, much larger. You may have noticed that this paper is extremely difficult to understand; we assure you that we also find it so. So why are we asking you to read it? Genetics has become extremely important in biomedical research, and it's important for psychiatrists, as physicians, to have knowledge about what advances have, and have not, been made, and how these may change the field of practice in the future. Consumer genetic testing is already attempting to make changes to medicine, and understanding how basic research in genetics is done can help you make decisions as new technologies arrive on the market.

The Psychiatric Genomics Consortium (PGS, https://www.med.unc.edu/pgc/) is a group of international investigators who work together without centralized funding. The PGS exists because very, very large samples (on the order of hundreds of thousands) are needed to study the genetics of psychiatric disorders. For this project, data from 39 sets of subjects was used. Detailed information about how the subjects were recruited and the data was analyzed is available in a "supplement" that you can download from the same website hosting the paper itself. We used the supplemental data file in writing this post-guide; it isn't necessary for you to read it, just be assured it is available for those who want to know more detail. Although all the subjects were white Europeans (for consistency of genetic background), they came from different countries and were enrolled using different types of assessments and criteria. They also were, presumably, genotyped using different technologies, though no information about this is provided. The analysis started with the raw genetic data from all the cohorts processing it using the same steps, or "pipeline." The chips used to determine this data use a variety of technologies but all of them return a value for a particular position on the genome with a specified certainty, either A, T, C or G. Usually only values with 95% certainty or better are used. Various chips will not include all the same genome locations, but because much of the genome is in linkage disequilibrium it is easy to guess, or impute, what other genome location values are if you know some of the values nearby. However, despite this, results are reported in terms of 'loci' in the genome that are more broad because of the statistical uncertainty that cannot be avoided entirely. The significant loci are small enough to be attached to a particular gene most of the time, though you will notice in Table 1 that one locus falls in a region between genes.

In this particular case, they used a Principal Component Analysis (PCA) to eliminate some of the data, only keeping variants for which the PCs correlated to the disease state. This method is also helpful for looking at patterns in the data that might be caused by other things, such as 'population stratification,' which basically means the sample is made up of subgroups who are more closely related to other members of the subgroup than the subgroups are to each other. In fact, the vast majority of the work done in this paper was to identify and minimize the effects of such confounding variables. This had to be done for each set of data, the discovery and replication sets, using the same methods, before the actual GWAS was performed.

While the text is very obscure, you should be able to have some insight into the table and figure, so we will now focus on how to read those. GWAS data are most typically presented in Manhattan plots (named because they are thought to look like the skyline of Manhattan), such as figure 1. The y axis is p values, and the line drawn horizontally represents the threshold for significance which is very, very high given the number of loci put into the analysis. The 'peaks' that cross the line are shown as the first column in table 1. The second column is the best guess as to which exact location in the genome (Single Nucleotide Polymorphism or SNP) is responsible for the significance, though recall because of the nature of the data the investigators cannot be 100% sure that these SNPs are the 'right' ones. The next two columns tell us which chromosome the loci are on, and the loci sizes, in base pairs. The next two contain the information, for the SNPs in column 2, of the bases that are found there, and then the percent of the total bases across the sample at that location that were the first base, so in the first row, SNP rs7544145 is a T 81% of the time.

Finally we get to the significance values, in this case for the first set of data, then the second, then for a meta-meta-analysis of both sets. Notice that the genes included in the table are chosen because they are the top genes for the meta-meta-analysis – all the p values are bolded – and only some of them were significant in the first data set alone. Apparently they used an uncorrected threshold of p = 0.05 for the replication data set, as all the values smaller than this are bolded. You can see that none of these would have passed a correct significance threshold in the replication. It seems to me that because priority was given to the combined data, the spirit of a primary analysis with a replication set is a little violated – the PGC previously set as their own standard that all publications must use the primary + replication formula so they set themselves up and then, seemingly try to wiggle out a bit here. Another way to look at the data that may yield a larger more global view of the gene variants enriched in bipolar is to calculate a polygenic risk score (PRS) which they mention doing in the text. The PRS accounted for 8% of the incidence of bipolar in this sample, which may sound puny if you don't know that the best we've been able to do for depression, for example, is about 2%.

The next step is to look at the genes in the list to see what proteins they code. If you are curious enough to do this yourself, you can search for genes at NCBI by pulling the menu to the left of the search bar to 'Gene' instead of 'PubMed.' The last few paragraphs of the paper (page 797) discuss what the authors think. Only a few of the genes on the list are obviously involved in neural processes or psychiatric function, like *GRIN2A*, which encodes a subunit of the Glutamate Receptor. An *a priori* way of learning more about a list of genes is via pathway analysis, which the authors also performed on their gene list; the top hits are insulin signaling and endocannabinoid signaling. Insulin receptors are present on some brain cells and insulin signaling may possibly be relevant, however because the pathway database is not tissue specific, it is also possible that this combination of genes is doing something else in the brains of patients. The authors don’t have much to say about endocannabinoids, but we may wonder if the fondness of many of our patients for THC containing compounds may be related to this finding, or if CBD based drugs may show some promise. So far, besides struggling to come up with consistent hits the biggest shortcoming of GWAS in psychiatry has been the failure of results to yield meaningful insights into the disease biology or treatment. Will these pathways prove useful? Only time will tell.

Pre-Guide (PGY2-4)

Raskind M.A. *et al*. 2013. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry*. 170: 1003-1010.

Reasons for choosing this article

- This article provides data on the use of prazosin in patients with PTSD. Many residents have clinical experience with this treatment, so it is important to understand what the data actually tell us.
- This conversation lets us have a bigger discussion about the gaps between the conduct of clinical trials and actual clinical practice.

Background

- Why do the authors think this is an important study?
- What are potential differences between military and civilian populations with PTSD?
- In your own words, what was the authors' hypothesis? In what ways is this hypothesis different from a hypothesis that prazosin will be effective treatment for PTSD?

Methods

- Who were the study participants? Do you consider the study inclusion and exclusion criteria to be broad or narrow? On what basis?
- The authors note that the member of the study team who adjusted medication dose was different from the person who performed/rated the assessments. Why split up these roles?
- What things were measured by the primary outcome measures? What does this tell us about how the authors defined response to medication?
- How was "responder" defined?

A technical point from the Background:

At the end of the Background section, the authors make the following statement: "We report the results of a pre-specified interim analysis that prompted the . . . Institutional Review Board to discontinue enrollment because of demonstrated efficacy." What does this mean? What is an Institutional Review Board (IRB)? Who are the members of the IRB? Under what circumstances would an IRB end a study?

Results

- What were the average doses of prazosin used in the study? How do these doses compare to your clinical experience with prazosin?

- Look at the results as presented in Figure 1. Footnote a states that the figures are "based on linear mixed effects models." What does this mean?
- Which symptom domains/measures were affected by prazosin treatment? Which were not affected? Are these results consistent with the biological rationale of using prazosin to target symptoms?
- What were the results in patients maintained on a stable SSRI dose during the study? What are the limitations on interpreting these results?

Discussion

- What do you take away from this study?
- Are these findings clinically important (in addition to being statistically significant)? Do you agree that the findings presented here are important enough for this study to be published in *American Journal of Psychiatry*, one of the top journals in our field?
- For which patients would you recommend prazosin? What symptoms would you expect it to decrease?
- The authors state that their "results cannot be extrapolated to persons with PTSD who do not recall trauma nightmares." What does this mean?

Post-Guide (PGY2-4)

Raskind M.A. *et al*. 2013. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry*. 170: 1003-1010.

Article Summary

This article is a fairly straightforward clinical trial of prazosin for sleep related problems, particularly nightmares, in patients with combat PTSD. Although there was fairly substantial evidence suggesting the prazosin is helpful for PTSD prior to this paper, it adds several important findings to the literature.

1) Previous studies dosed prazosin only at night and were unable to determine whether there was overall (as opposed to just sleep related) improvement in PTSD symptoms.
2) Little is known about the use of prazosin with other psychotropics
3) This study enrolled OIF/OEF veterans still on active duty with recent and perhaps future trauma. In contrast, prior studies of the efficacy of prazosin were conducted in Vietnam veterans years after exposure.

Subjects were enrolled and randomly assigned to receive either placebo or prazosin dosed twice a day, with a much smaller dose in the morning. Blinded medications were adjusted via the number of pills taken without revealing the underlying dose in mg. Women received a different dosing schedule than men - a very, very unusual study design element. Dose titration took place over 6 weeks; there was no assessment of symptoms until week 7. A rater blinded to most information about each subject gave the Clinician Administered PTSD scale (CAPS) as the primary outcome. Sub-scales of the CAPS (the item assessing nightmares in particular), the Hamilton Depression Rating Scale, and a scale to assess sleep were also collected. Side effects were assessed at each visit (not just at the visits that collected symptom data), presumably with a form that ensured consistent info was collected about each (usually what, when, how bad, when resolved, and if interventions were made). The assessment period was 15 weeks long.

The results are notable for their consistency. Prazosin beat placebo in nearly every measure examined: nightmares, sleep overall, and total CAPS score. Secondary analysis suggested this improvement was likely driven by the changes in sleep symptoms. Perhaps surprisingly, an analysis of the subset of patients taking both prazosin and an SSRI demonstrated that SRRIs seemed to inhibit improvement from taking prazosin.

Comments

This study is important because it has high clinical utility in suggesting how prazosin should be used in PTSD. A big part of the reason we chose it is because we've observed that low (likely sub-therapeutic) doses are often tried due to fear of side effects, and that therefore clinicians may conclude that prazosin is less effective than studies do. This study suggests that in healthy adults fairly high doses are tolerated well.

The finding that sleep symptoms are the primary target of prazosin is also interesting. Because the study did not complete enrollment (see Technical Point below), we don't know whether change in other symptom domains in the CAPS would have eventually separated statistically. Furthermore, we don't know how to interpret the finding that patients on SSRIs did not respond as well to prazosin. It could be due to the sleep disrupting effects of SSRIs, or it might because patients with SSRI treatment were somehow different from patients not treated with SSRIs. An unknown difference might have caused patients on SSRIs to respond less well. In particular, if those on SSRI were more sick at baseline they may have not responded as well; however they do not indicate that baseline severity of PTSD or depression symptoms was included as a term in the analysis model. The relatively small number of patients on SSRIs included in the study likely means that this analysis was underpowered, preventing the authors from answering many of these questions.

Technical Point

Research ethics can easily be overlooked when reading the literature but is of critical importance. The main method of ensuring research with human subjects is conducted ethically is via institutional review. Typically a university, hospital, company or other entity will operated a number of Institutional Review Boards (IRBs) to review proposals for humans subjects research.

IRBs are tasked with applying principles of research ethics to specific studies. These principles are set out officially in the US with something called the "common rule" which itself is based on various historical statements on research ethics including the results of the investigation into war crimes committed by Nazi scientists in World War II (the "Nuremberg Code") and a major convention in the 1970s (the "Belmont Report"). If you ever participate in a study you will learn about these in the required training. The essential principles involve education of subjects so they can voluntarily participate, balancing risks and benefits of research, and wider concerns about how research should function within the health care system and society at large.

In this trial, the IRB at the study site made a recommendation to stop the study. This was based on an interim analysis (an analysis planned after a certain percentage of enrollment is complete, or, sometimes, after a set time period). The purpose of interim analysis is to check to be sure the risks and benefits of the study are not so different from predicted that there is reason to reevaluate the ethical status of the protocol. The prototypical example is that subjects may be dying unexpectedly because of the treatment – say a toxic chemotherapy – and continuing to enroll subjects exposes them to unethical risk. In this case, however, the study was ended for *benefit*. This means that all the primary measures separated from placebo early – before the enrollment was complete – so the IRB decided there was no reason to keep exposing subjects to placebo. They asked the investigators to stop the study and offer treatment to all of the participants. We won't go into whether or not this is the "right" decision, instead our purpose is to make you aware that researchers are subject to ethical review which sometimes has a profound impact. The IRB may also ask for the protocol to be modified, the kind of patients recruited to change, or even block a study completely. They also handle complaints from research subjects.

Pre-Guide (Intern)

Raskind M.A. *et al*. 2013. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry*. 170: 1003-1010.

Accompanying design and statistics article

Pocock SJ and Stone GW. 2016. The Primary Outcome is Positive—Is that Good Enough? *NEJM* 375(10): 971-979.

Reasons for choosing this article

- This article provides data on the use of prazosin in patients with PTSD. Many residents have clinical experience with this treatment, so it is important to understand what the data actually tell us.
- This article lets us think about the interpretation of studies when the primary outcome is positive.

Background

- Why do the authors think this is an important study?
- What are potential differences between military and civilian populations with PTSD?
- In your own words, what was the authors' hypothesis? In what ways is this hypothesis different from a hypothesis that prazosin will be effective treatment for PTSD?

Methods

- Who were the study participants? Do you consider the study inclusion and exclusion criteria to be broad or narrow? On what basis?
- The authors note that the member of the study team who adjusted medication dose was different from the person who performed/rated the assessments. Why split up these roles?
- What things were measured by the primary outcome measures? What does this tell us about how the authors defined response to medication?
- How was "responder" defined?

A technical point from the Background:

At the end of the Background section, the authors make the following statement: "We report the results of a pre-specified interim analysis that prompted the . . . Institutional Review Board to discontinue enrollment because of demonstrated efficacy." What does this mean?

Results

- What were the average doses of prazosin used in the study?
- Look at the results as presented in Figure 1. Footnote a states that the figures are “based on linear mixed effects models.” What does this mean? Are the raw data presented in this figure, or have they been analyzed in some way?
- Which symptom domains/measures were affected by prazosin treatment? Which were not affected? Are these results consistent with the biological rationale of using prazosin to target symptoms?
- What were the results in patients maintained on a stable SSRI dose during the study? What are the limitations on interpreting these results?

Discussion

- What do you take away from this study?
- In the accompanying article, Pocock and Stone raise several “key questions” to be considered in studies in which the primary outcome is positive and note that “concerns” may emerge when a reader tries to answer these questions. How do you answer those questions for this study? What concerns do you have about this study?
- The authors state that their “results cannot be extrapolated to persons with PTSD who do not recall trauma nightmares.” What does this mean? Why do they make this statement?

Post-Guide (Intern)

Raskind M.A. *et al*. 2013. A Trial of Prazosin for Combat Trauma PTSD with Nightmares in Active-Duty Soldiers Returned from Iraq and Afghanistan. *Am J Psychiatry*. 170: 1003-1010.

Article Summary

This article is a fairly straightforward clinical trial of prazosin for sleep related problems, particularly nightmares, in patients with combat PTSD. Although there was fairly substantial evidence suggesting the prazosin is helpful for PTSD prior to this paper, it adds several important findings to the literature.

4) Previous studies dosed prazosin only at night and were unable to determine whether there was overall (as opposed to just sleep related) improvement in PTSD symptoms.
5) Little is known about the use of prazosin with other psychotropics
6) This study enrolled OIF/OEF veterans still on active duty with recent and perhaps future trauma. In contrast, prior studies of the efficacy of prazosin were conducted in Vietnam veterans years after exposure.

Subjects were enrolled and randomly assigned to receive either placebo or prazosin dosed twice a day, with a much smaller dose in the morning. Blinded medications were adjusted via the number of pills taken without revealing the underlying dose in mg. Women received a different dosing schedule than men - a very, very unusual study design element. Dose titration took place over 6 weeks; there was no assessment of symptoms until week 7. A rater blinded to most information about each subject gave the Clinician Administered PTSD scale (CAPS) as the primary outcome. Sub-scales of the CAPS (the item assessing nightmares in particular), the Hamilton Depression Rating Scale, and a scale to assess sleep were also collected. Side effects were assessed at each visit (not just at the visits that collected symptom data), presumably with a form that ensured consistent info was collected about each (usually what, when, how bad, when resolved, and if interventions were made). The assessment period was 15 weeks long.

The results are notable for their consistency. Prazosin beat placebo in nearly every measure examined: nightmares, sleep overall, and total CAPS score. Secondary analysis suggested this improvement was likely driven by the changes in sleep symptoms. Perhaps surprisingly, an analysis of the subset of patients taking both prazosin and an SSRI demonstrated that SRRIs seemed to inhibit improvement from taking prazosin.

Comments

This study is important because it has high clinical utility in suggesting how prazosin should be used in PTSD. A big part of the reason we chose it is because we've observed that low (likely sub-therapeutic) doses are often tried due to fear of side effects, and that therefore clinicians may conclude that prazosin is less effective than studies do. This study suggests that in healthy adults fairly high doses are tolerated well.

The finding that sleep symptoms are the primary target of prazosin is also interesting. Because the study did not complete enrollment (see Technical Point below), we don't know whether change in other symptom domains in the CAPS would have eventually separated statistically. Furthermore, we don't know how to interpret the finding that patients on SSRIs did not respond as well to prazosin. It could be due to the sleep disrupting effects of SSRIs, or it might because patients with SSRI treatment were somehow different from patients not treated with SSRIs. An unknown difference might have caused patients on SSRIs to respond less well. In particular, if those on SSRI were more sick at baseline they may have not responded as well; however they do not indicate that baseline severity of PTSD or depression symptoms was included as a term in the analysis model. The relatively small number of patients on SSRIs included in the study likely means that this analysis was underpowered, preventing the authors from answering many of these questions.

Technical Point

Research ethics can easily be overlooked when reading the literature but is of critical importance. The main method of ensuring research with human subjects is conducted ethically is via institutional review. Typically a university, hospital, company or other entity will operated a number of Institutional Review Boards (IRBs) to review proposals for humans subjects research.

IRBs are tasked with applying principles of research ethics to specific studies. These principles are set out officially in the US with something called the "common rule" which itself is based on various historical statements on research ethics including the results of the investigation into war crimes committed by Nazi scientists in World War II (the "Nuremberg Code") and a major convention in the 1970s (the "Belmont Report"). If you ever participate in a study you will learn about these in the required training. The essential principles involve education of subjects so they can voluntarily participate, balancing risks and benefits of research, and wider concerns about how research should function within the health care system and society at large.

In this trial, the IRB at the study site made a recommendation to stop the study. This was based on an interim analysis (an analysis planned after a certain percentage of enrollment is complete, or, sometimes, after a set time period). The purpose of interim analysis is to check to be sure the risks and benefits of the study are not so different from predicted that there is reason to reevaluate the ethical status of the protocol. The prototypical example is that subjects may be dying unexpectedly because of the treatment – say a toxic chemotherapy – and continuing to enroll subjects exposes them to unethical risk. In this case, however, the study was ended for *benefit*. This means that all the primary measures separated from placebo early – before the enrollment was complete – so the IRB decided there was no reason to keep exposing subjects to placebo. They asked the investigators to stop the study and offer treatment to all of the participants. We won't go into whether or not this is the "right" decision, instead our purpose is to make you aware that researchers are subject to ethical review which sometimes has a profound impact. The IRB may also ask for the protocol to be modified, the kind of patients recruited to change, or even block a study completely. They also handle complaints from research subjects.

Accompanying design and statistics article

This article addresses our tendency to think of studies in literature as demonstrating whether or not a medication works, reminding us to pause and carefully examine the methods and results in any study in which the primary outcome is positive. The authors specifically identify a useful set of questions to consider with any trial with positive primary outcome results:

- Does a statistically significant *p* value provide strong enough evidence?
- What is the magnitude of the treatment benefit?
- Is the primary outcome clinically important (and internally consistent)?
- Are secondary outcomes supportive?
- Are the principal findings consistent across important subgroups?
- Is the trial large enough to be convincing?
- Was the trial stopped early?
- Do concerns about safety counterbalance positive efficacy?
- Is the efficacy-safety balance patient specific?
- Are there flaws in the trial design and conduct?
- Do the findings apply to my patients?

Keep this list handy as you read clinical trials throughout residency (and your career after residency).

Pre-Guide (PGY2-4)

GS Sachs *et al*. (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine* 356(17):1711-1722.

Reasons for choosing this article

- This manuscript describes the primary outcomes of the STEP-BD trial, which is one of a handful of major effectiveness trials in psychiatry.
- The article addresses a major clinical question: what is the role of antidepressants in the treatment of bipolar depression?

Background

- What is an effectiveness trial? How does it differ from an efficacy study?
- What do the authors give as the reasons for conducting this study? How do they believe it differs from previous medication trials for bipolar depression?
- What do you think was the hypothesis of study?

Methods

- Who were the study participants?
- Do you agree with the medications that were considered "mood stabilizers"?
- What rationale do the authors give for using paroxetine and bupropion as the antidepressants? Do you agree with the choice to use these two medications given the aim of the study?
- How long did study participation last? What determined how long participants remained in study and how often study visits occurred?
- How were the study outcomes defined?

A technical point from the Methods:

The authors describe using an “equipoise-stratified randomization method.” In your own words, what do they mean by this?

Results

- What are the major findings of the study? What effect did the addition of an antidepressant have to a mood stabilizer for the treatment of bipolar depression? Was there harm associated with antidepressant treatment?
- Did outcomes differ between patients with bipolar I vs bipolar II?
- What are the differences between each of the outcomes listed in Table 4? Why do you think the authors included each of these outcomes?

Discussion

- What do you take away from this study?
- The authors note that the study participants “were already receiving clinical treatment at participating sites and . . . continued care with their usual provider.” Is this a strength or weakness of this study?
- In the Background, the authors state their goals included recruitment of "a representative group of patients" and measure "clinically meaningful outcomes." Did they meet these goals? In other words, did they meet the goal of completing an effectiveness trial?
- On page 1720, the authors refer to a meta-analysis by Gijsman et al that examined the efficacy of antidepressants in the treatment of bipolar depression. This meta-analysis had results that differed from STEP-BD. Which results do you think are more reliable?
- How often do you encounter patients with bipolar disorder who are treated with both a mood stabilizer and an antidepressant? Are there clinical situations in which you would treatment a patient with bipolar depression with an antidepressant?

Post-Guide (PGY2-4)

GS Sachs *et al*. (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine* 356(17):1711-1722.

Take Home Summary

This article describes the results of a large (well, for a bipolar trial, n=366) effectiveness trial comparing mood stabilizer (“any FDA-approved antimanic agent”) alone to mood stabilizer plus an antidepressant (bupropion or paroxetine) in the treatment of bipolar depression. The study attempted to answer two questions: (1) does the addition of an antidepressant improve recovery from bipolar depression and (2) does the addition of an antidepressant increase the rate of switching from depression to mania? These questions are important because bipolar depression has, perhaps, the least positive evidence for any effective treatment. Today, several atypical antipsychotics have FDA approval for the treatment of bipolar depression, but given their unfavorable side effect profile compared to antidepressants, it remains common to add an antidepressant to a mood stabilizer in patients who complain of ongoing depressive episodes in bipolar disorder. Fifteen years ago when this study was designed, it was an even more critical gap in evidence based care for one of our toughest diseases.

Study participants were adults (age >18) with a current DSM-IV depressive episode associated with bipolar I or bipolar II disorder. They were recruited from 22 sites around the U.S. and were followed for up to 26 weeks. Participants got open-label, standard of care treatment with a mood stabilizer that was titrated based on clinical symptoms and were randomized in a double-blind fashion to concurrent treatment with bupropion, paroxetine, or placebo. Typically for effective studies, the primary outcome was “durable recovery,” which was defined as euthymia for at least 8 consecutive weeks. Other outcome measures were transient recovery (1-7 weeks of euthymia), treatment-emergent affective switch, no response (no euthymia lasting at least 1 week with 16 weeks of treatment), and treatment-effectiveness response (50% improvement in depressive symptoms without meeting criteria for mania or hypomania).

Overall, there were no significant differences between groups on any of the outcomes—the rate of improvement in depressive symptoms was the same whether or not patients were treated with an antidepressant. In fact the data came close to favoring the combination of mood stabilizer + placebo over mood stabilizer + antidepressant. Fortunately, there was also no increased risk of switch to mania in the patients treated with an antidepressant. Thus the major take home from this study was that antidepressants are ineffective in treating bipolar depression. Are you surprised by this finding? These results are now ten years old – unfortunately the relatively desperate need for treatments for bipolar depression held back changes in practice. Now that more evidence supports atypicals as effective treatment in these patients, psychiatrists have a viable alternative for evidence-based care.

Technical Point

STEP-BD is described as having “equipoise stratified” randomization. The term **equipoise** means that two clinical treatment options are both standard of care. It reflects the natural historical development of the practice of medicine and the social nature of the standard of care, in which new physicians are trained to do things based on local or regional traditions as well as on evidence. This acknowledges that clinicians and patients have preferences but that social consensus is not the same thing as evidence. In order to be considered in equipoise, treatments should be accepted by most and preferred by at least some doctors.

When treatments are in equipoise they are considered to be ethically viable options for randomization. This must take into account risks and side effects as well as efficacy. If a treatment may be worse than placebo (that is, cause harm without efficacy) then a placebo group is typically considered acceptable, but in cases where delayed treatment is not standard of care, placebo would not be considered in equipoise with a proposed therapy. In STEP-BD, for example, it would not be acceptable to randomize to antidepressant or placebo alone because no treatment is not standard of care for bipolar disorder. However, augmentation with an antidepressant, and mood stabilizer alone are (or were) both standard of care options for bipolar disorder and can be considered to be in equipoise because the risks of antidepressants (especially risk of mania) may be greater than the possible benefits.

STAR*D was designed to take this principle one step further – noting that patients who had a strong preference for a particular form of treatment, medication over psychotherapy say, would be less likely to enroll in a study randomizing across those options. They opted to include patient preference in the randomization scheme. Since whether a person is willing to engage with a particular treatment is an important factor in effectiveness (as opposed to efficacy) many studies following STAR*D adopted this methodology to increase enrollment and better replicated real world treatment conditions. So, STEP-BD subjects could choose which treatment arms they were willing to consider and then be randomized among only that subset as long as the subset were in equipoise with each other.

Want to learn more on this topic of equipoise in research? Check out this two page review:
AJ London. (2017) Equipoise in Research: Integrating Ethics and Science in Human Research. *JAMA* 317(5): 525-526.

Pre-Guide (Intern)

GS Sachs *et al*. (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine* 356(17):1711-1722.

Detry and Lewis (2014). The Intention-to-Treat Principle: How to Assess the True Effect of Choosing a Medical Treatment. *JAMA* 312(1):85-86.

Reasons for choosing this article

- This manuscript describes the primary outcomes of the STEP-BD trial, which is one of a handful of major effectiveness trials in psychiatry.
- The article addresses a major clinical question: what is the role of antidepressants in the treatment of bipolar depression?
- Nearly every clinical trial we read is analyzed according to the intent-to-treat principle. It's important to understand what this is and how it effects our interpretation of study outcomes.

Background

- What do the authors give as the reasons for conducting this study? How do they believe it differs from previous medication trials for bipolar depression?
- What do you think was the hypothesis of study?

Methods

- Who were the study participants?
- Do you agree with the medications that were considered "mood stabilizers"?
- What rationale do the authors give for using paroxetine and bupropion as the antidepressants? Do you agree with the choice to use these two medications given the aim of the study?
- How were the study outcomes defined?

Results

- What are the major findings of the study? What effect did the addition of an antidepressant have to a mood stabilizer for the treatment of bipolar depression? Was there harm associated with antidepressant treatment?
- Did outcomes differ between patients with bipolar I vs bipolar II?
- What are the differences between each of the outcomes listed in Table 4? Why do you think the authors included each of these outcomes?

- Were the outcomes assessed according to the intent-to-treat principle? Is this the right approach to take in this study? How might analysis utilizing a per-protocol approach affect our understanding of the data?

Discussion

- What do you take away from this study?
- In the Background, the authors state their goals included recruitment of "a representative group of patients" and measure "clinically meaningful outcomes." Did they meet these goals?
- Do you agree with the limitations identified by the authors? Are there any other limitations you would add?
- In that intention-to-treat article, Detry and Lewis make the following statement: “A characteristic of the ITT principle is that poor treatment adherence may result in lower estimates of treatment efficacy and a loss of study power. However, these estimates are clinically relevant because real-world effectiveness is limited by the ability of patients and clinicians to adhere to a treatment.” What is the relevance of this statement to STEP-BD and the other trials we’ve discussed this year?
- How often do you encounter patients with bipolar disorder who are treated with both a mood stabilizer and an antidepressant? Are there clinical situations in which you would treatment a patient with bipolar depression with an antidepressant?

Article

GS Sachs *et al*. (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine* 356(17):1711-1722.

Detry and Lewis (2014). The Intention-to-Treat Principle: How to Assess the True Effect of Choosing a Medical Treatment. *JAMA* 312(1):85-86.

Take Home Summary

This article describes the results of a large (n=366) effectiveness trial comparing the efficacy of any mood stabilizer ("any FDA-approved antimanic agent") alone compared to any mood stabilizer plus an antidepressant (bupropion or paroxetine) in the treatment of bipolar depression. The study attempted to answer two questions: (1) does the addition of an antidepressant improve recovery from bipolar depression and (2) does the addition of an antidepressant increase the rate of switching from depression to mania? Study participants were adults (age >18) with a current DSM-IV depressive episode associated with bipolar I or bipolar II disorder. They were recruited and seen at 22 sites around the U.S. and were followed for up to 26 weeks. Participants were treated openly with a mood stabilizer that was titrated based on clinical symptoms and were randomized in a double-blind fashion to concurrent treatment with bupropion, paroxetine, or placebo. The primary outcome was "durable recovery," which was defined as euthymia for at least 8 consecutive weeks. Other outcome measures were transient recovery (1-7 weeks of euthymia), treatment-emergent affective switch, no response (no euthymia lasting at least 1 week with 16 weeks of treatment), and treatment-effectiveness response (50% improvement in symptoms without meeting criteria for mania or hypomania). Overall, there were no significant differences between groups on any of the effectiveness outcomes—the rate of improvement in depressive symptoms was the same whether or not patients were treated with an antidepressant. Trends in the outcomes, however, favored the combination of mood stabilizer + placebo over mood stabilizer + antidepressant. Additionally, there was no increased risk of affective switch in the patients treated with an antidepressant in addition to a mood stabilizer compared to those treated with a mood stabilizer and a placebo. The authors report the major conclusion from this paper as "no evidence that treatment with a mood stabilizer and an antidepressant confers a benefit over treatment with a mood stabilizer alone." Others, however, have emphasized the finding of no increased risk of affective switch when an antidepressant is used along with a mood stabilizer.

The intention-to-treat principle is a guiding principle for the analysis of clinical trial data. According to this principle, data from all randomized study participants are analyzed by the group to which to the participant was assigned. In other words, a participant assigned to receive treatment A is included in the analysis of the outcomes regardless of how much of treatment A the participant actually received. An intention-to-treat analysis can be contrasted against a per-protocol analysis. In a per-protocol analysis, only the participants were who adherent to the treatment and study visits are included in the analysis. It may seem like the per-protocol analysis is the logical way to analyze data, in that (under this approach), you measure the effect of an intervention in the people who go the intervention the way the study intended. The problem with this approach, however, is that study drop out/people not being adherent to the medication is likely not to be random. For example, study participants might stop taking treatment A due to a side effect; if these participants are removed from the analysis, you will overestimate the benefit of treatment A. Additionally, real world patients rarely are perfectly adherent to interventions, so the per protocol analysis will likely overestimate the actual clinical benefit of an intervention. As Detry and Lewis note, there are situations in which a per-protocol analysis is appropriate, like an early efficacy (Phase 2) trial. There are also times in which the authors of a study will present both the intention-to-treat analysis and the per-protocol analysis, particularly if the study results differ between the two. A recent example of this is: JD Lee *et al*. (2018) Comparative effectiveness of extended-release naltrexone versus buprenorphine-naloxone for opioid relapse prevention (X:BOT): a multicentre, open-label, randomised controlled trial. *Lancet* 391(10118): 309-318.

Pre-Guide (PGY2-4)

Schneider LS *et al*. 2006. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *New England Journal of Medicine* 355(15):1525-1538.

Reasons for choosing this article

- This article reports the primary outcomes of phase I of the CATIE AD trial, a major effectiveness trial for antipsychotics in Alzheimer's disease. It is a landmark study.
- Given the FDA black box warning regarding the use of antipsychotics in patients with dementia, it is important to understand the risks and benefits of these medications in this population.

Background

- What has been your clinical experience with patients with dementia who have hallucinations and/or delusions? What medications have been used to target these symptoms?
- What was the rationale for the study? What was the hypothesis?
- Was CATIE AD designed to be an efficacy or effectiveness trial?

Methods

- What patients were included in the study? How was the diagnosis of Alzheimer's disease determined? What about hallucinations/delusions/aggression?
- Why did the authors require that every study participant had to have a "study partner or caregiver" participate in the assessments?
- How did the authors of CATIE AD design the pills that were used in the study?
- Why did the authors choose discontinuation of treatment as the primary outcome? Do you agree with this choice?
- Do you think the study had appropriate power?

A technical point from the Methods:

Towards the end of the statistical analysis section, the authors discuss performing testing "equivalence" and using a "one-sided test with a P value of less that 0.025" for some comparisons. What does this mean? What is equivalence testing? What is the difference between a two-sided and a one-sided *p* value?

Results

- Looking at Table 1, what were the characteristics of the patients in the trial? What was the typical cognitive function/disease severity in study participants?

- What are the results with regards to the primary outcome (discontinuation for any cause)? Did any drug stand out? How did placebo perform? Did this change based on the outcome measure (compare Fig 2 A vs B vs C)?
- How do the results of this trial (CATIE AD) compare to the CATIE schizophrenia trial?
- Why did the authors include information on participant caregivers in the results?
- Do the reported side effects match with your clinical experience? What do make of the changes in weight and prolactin levels over the 12 weeks of the trial—are these concerning?

Discussion

- What do you take away from this study?
- How do these results affect your determination of the risk/benefit profile for the use of antipsychotics in patients with Alzheimer's disease?
- The authors suggest that the study physicians may have discontinued patients on medication in phase 1 quickly in order to put the participants into phase 2. What do you make of this suggestion? Do you think this explains any of the study findings?
- The authors state "the key enrollment criteria—the physician's assessment that an antipsychotic drug was the appropriate therapy—helped to ensure clinical equipoise." What do they mean by this?
- The FDA has stated "antipsychotics are not indicated for the treatment of dementia-related psychosis" and the study authors state "our findings suggest that there is no large clinical benefit of treatment with atypical antipsychotic medications." Do you agree? Why do you think these medications continue to be used clinically?

Post-Guide (PGY2-4)

Schneider LS *et al*. 2006. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *New England Journal of Medicine* 355(15):1525-1538.

Article Summary

This is the main report from the CATIE-AD trial, a sister study to CATIE, examining the use of atypical antipsychotics (AAPs) for behavioral symptom management in patients with Alzheimer's Disease. Much like CATIE, this is an effectiveness trial using time to discontinuation and change on the Clinical Global Impression scale as outcome measures. Subjects met criteria for Alzheimer's Dementia and were experiencing psychotic symptoms or behavioral problems judged to be of moderate severity and regular occurrence. Subjects were randomized to one of three blinded antipsychotic drugs (olanzapine, quetiapine, risperidone) or placebo for two weeks, during which they were required to stay on the study drug, with dose adjustment. Subjects then entered a follow up phase on which continuation, dose, side effects and CGI scores were tracked, with a primary endpoint at 12 weeks.

Overall, the majority of patients discontinued drugs by week 12, over 60%, with no significant differences. However, both risperidone and olanzapine were maintained longer than quetiapine and placebo when considering stopping for lack of efficacy alone. Few subjects stopped taking the drug for intolerance, but overall the statistics here favored placebo. The authors then compared response on the CGI (defined as at least minimally improved), but did not find differences between the treatment arms.

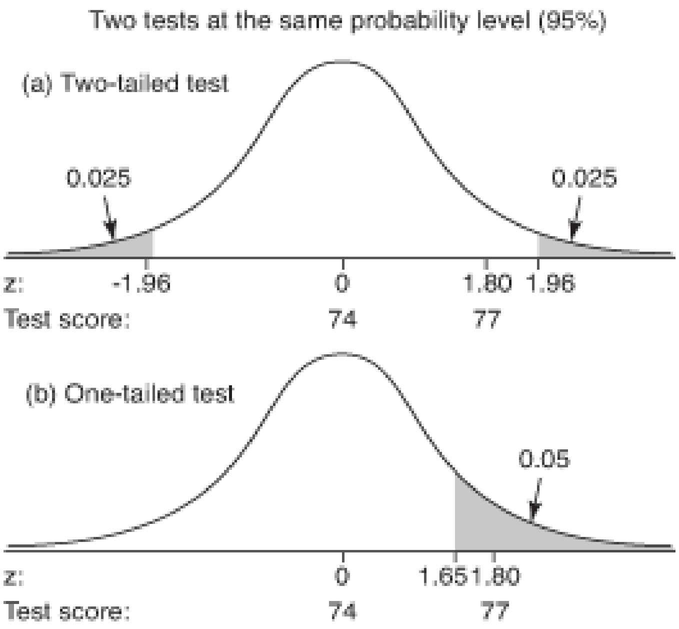
Comments

CATIE-AD was a response to the addition of federal warnings on the use of antipsychotic drugs in the elderly. Because use in this population of elderly persons was, and is, common, the study sought to validate the risk/benefit for AAPs.

Overall the study found insufficient evidence to recommend the use of AAPs in patients with dementia. These results strongly suggest that these drugs should not be used in the elderly except in patients who are carrying an indicated diagnosis such as schizophrenia into old age, and for short term as needed use for delirium. It's important to note that because CATIE-AD used broad general criteria for response as opposed to specific measures of symptoms, and involved caregivers in the assessments, it can't effectively be argued that there are secondary benefits to AAP treatment in dementia, such as making patients more cooperative with caregivers. The fact that CATIE-AD failed to have a significant impact on clinical practice is troubling.

Technical Point

The CATIE-AD authors did something unusual in performing a test of equivalence (or non-superiority). In non-superiority or equivalence testing the null hypothesis and study hypothesis are reversed, and you try to demonstrate that within some clinical margin two treatments are equally good. This type of testing is often done when a new expensive drug us approved and there is a question of



whether the cost is justifiable. Here the authors are simply interested in comparing the drugs used in the trial.

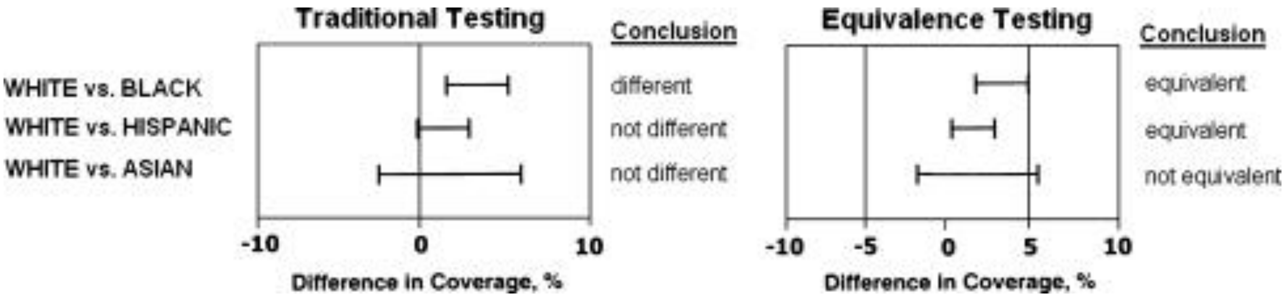
Most analysis use the two tailed t-test which leaves open the idea that either group could be superior (that is that the difference between groups could be in either direction) On the other hand, one tailed tests assume that the direction of a possible difference is known, or that a difference in only one direction is meaningful.

The figure above shows the difference between one and two tailed tests mathematically. The curves shown above are NOT study data but instead the curve of Z-scores that represent the probability function of the outcome variable of interest. You can see in the figure that the region of significance is split in a two tailed test compared to a one-tailed test.

But what does this have to do with equivalence? To test for equivalence you first set a range that defines the difference between groups that would be acceptable. For example, you could decide that a difference of less than 5 points on a scale between groups is not clinically significant (perhaps because in that range there is no difference in long term prognosis or functional capacity). Contrast this to a typical comparison using a two tailed test (see the figure below from Walker and Nowaki et al (2011)). Usually the null hypothesis is represented by one line where "the difference between groups is zero" and if your confidence interval includes this line, the results are not significant.

In equivalence testing you have **two** lines of interest and you must perform two separate one tailed t-tests. One test establishes the confidence interval is **below the maximum**, and the other test that it is **above the minimum**, of the range. Because the confidence interval must fall on a particular side of each line, one tailed tests are appropriate.

.



Pre-Guide (Intern)

Schneider LS *et al*. 2006. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *New England Journal of Medicine* 355(15):1525-1538.

Tolles J and Lewis R. 2016. Time to Event Analysis. *JAMA* 315 (10): 1046-1047.

Reasons for choosing this article

- This article reports the primary outcomes of phase I of the CATIE AD trial, a major effectiveness trial for antipsychotics in Alzheimer's disease. It is a landmark study.
- Given the FDA black box warning regarding the use of antipsychotics in patients with dementia, it is important to understand the risks and benefits of these medications in this population

Background

- What has been your clinical experience with patients with dementia who have hallucinations and/or delusions? What medications have been used to target these symptoms?
- What was the rationale for the study? What was the hypothesis?
- Was CATIE AD designed to be an efficacy or effectiveness trial?

Methods

- What patients were included in the study? How was the diagnosis of Alzheimer's disease determined? What about hallucinations/delusions/aggression?
- Why did the authors require that every study participant had to have a "study partner or caregiver" participate in the assessments?
- How did the authors of CATIE AD design the pills that were used in the study?
- Why did the authors choose discontinuation of treatment as the primary outcome? Do you agree with this choice?
- Do you think the study had appropriate power?

A technical point from the Methods:

What is a time-to-event analysis? How did the authors use time-to-event analyses? Was this an appropriate analysis for these data?

Results

- Looking at Table 1, what were the characteristics of the patients in the trial? What was the typical cognitive function/disease severity in study participants?

- What are the results with regards to the primary outcome (discontinuation for any cause)? Did any drug stand out? How did placebo perform? Did this change based on the outcome measure (compare Fig 2 A vs B vs C)?
- How do the results of this trial (CATIE AD) compare to the CATIE schizophrenia trial?
- Why did the authors include information on participant caregivers in the results?
- Do the reported side effects match with your clinical experience? What do you make of the changes in weight and prolactin levels over the 12 weeks of the trial—are these concerning?

Discussion

- What do you take away from this study?
- How do these results affect your determination of the risk/benefit profile for the use of antipsychotics in patients with Alzheimer's disease?
- The authors state "the key enrollment criteria—the physician's assessment that an antipsychotic drug was the appropriate therapy—helped to ensure clinical equipoise." What do they mean by this?
- The FDA has stated "antipsychotics are not indicated for the treatment of dementia-related psychosis" and the study authors state "our findings suggest that there is no large clinical benefit of treatment with atypical antipsychotic medications." Do you agree? Why do you think these medications continue to be used clinically?

Post-Guide (Intern)

Schneider LS *et al*. 2006. Effectiveness of Atypical Antipsychotic Drugs in Patients with Alzheimer's Disease. *New England Journal of Medicine* 355(15):1525-1538.

Tolles J and Lewis R. 2016. Time to Event Analysis. *JAMA* 315 (10): 1046-1047.

Article Summary

This is the main report from the CATIE-AD trial, a sister study to CATIE, examining the use of atypical antipsychotics for behavioral symptom management in patients with Alzheimer's disease. Much like CATIE, this is an effectiveness trial using time to discontinuation as the primary outcome and change on the Clinical Global Impression scale as a secondary measure. Study participants met criteria for dementia of the Alzheimer's type and were experiencing psychotic symptoms or behavioral problems judged to be functionally impairing and were occurring routinely. Participants were randomized to one of three blinded antipsychotic drugs (olanzapine, quetiapine, risperidone) or placebo and were required to remain on the study medication for at least 2 weeks, with dose adjustment per clinical judgment. Continuation, dose, side effects and CGI scores were tracked, with a primary endpoint at 12 weeks. The study hypothesis was that all of the medications would be perform better than placebo and that no medication would be inferior to any other medication.

The average time to discontinuation ranged from 5-8 weeks, with no significant differences between any of the groups. The majority (over 60%) of patients discontinued drugs by week 12. However, both risperidone and olanzapine were maintained longer than quetiapine and placebo when considering stopping for lack of efficacy alone. Few subjects stopped taking the drug for intolerance, but overall the statistics here favored placebo. There were no differences between groups on treatment response on the CGI (defined as at least minimally improved).

Comments

Evidence that the use of antipsychotics in elderly patients with dementia is associated with an increased risk of death emerged while the CATIE-AD trial was underway, and the study authors clearly had this in mind while writing this manuscript. Because use in this population of elderly persons was, and is, common, the study sought to validate the risk/benefit for atypical antipsychotics.

Overall the study found insufficient evidence to recommend the use of atypical antipsychotics in patients with dementia. These results strongly suggest that these drugs should not be used in the elderly except in patients who are carrying an indicated diagnosis such as schizophrenia into old age and for short term as needed use for delirium. It has been argued in the past that antipsychotics are of benefit in this population because the medications help patients be more cooperative with their caregivers. The evidence in the CATIE AD trial suggest this is not true, as CATIE-AD used broad general criteria for response as opposed to specific measures of symptoms and involved caregivers in the assessments. The fact that CATIE-AD failed to have a significant impact on clinical practice is troubling.

Technical Point

In CATIE-AD, the authors used a time-to-event analysis to estimate the time to discontinuation of medications in the intent-to-treat population. Time to discontinuation of a medication encompasses many variables of interest and, like CATIE, exemplifies a real-world clinical outcome of interest. Using a Kaplan-Meier survival curve estimates how many patients are still taking a medication at a particular point in time (or over a specific time interval, to be more precise). As discussed in the article by Tolles and Lewis, the advantage of using a time-to-event analysis instead of a hazard function analysis is that the time-to-event analysis includes participants who have the event as well as those who do not, while a hazard function includes only those with the event. To be more specific, we know that 80 of the 100 participants were assigned to olanzapine discontinued this medication in phase 1 while 20 remained on the medication. The hazard function analyzes only the 80 patients who discontinued, while the time-to-event analysis (the survival curve) includes all 100. The authors calculated the hazard ratio for each antipsychotic in comparison to placebo and with each other. This value estimates the risk of discontinuation of the medication over time due to both 1) lack of efficacy and (2) intolerability, adverse effects or death.

Pre-Guide

Skoglund, C *et al*. 2015. Attention-Deficit/Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biological Psychiatry* 77:880-886.

Reasons for choosing this article

- This study examines the concern that the use of stimulant medications for treatment of ADHD predisposes people to the development of substance use disorders (SUD).
- This study uses large population databases to perform a case-control study, and it is important to understand the advantages and disadvantages of this study design.

Background

- Prior to this study, what was known about the co-occurrence of ADHD and SUD? What is your clinical experience with this group of patients?
- Do you think the question examined in the study is an important one? Why or why not?
- What do you think was the authors' hypothesis? What was your expectation regarding the outcomes?

Methods

- Where did the data used in the study come from?
- Which groups did the authors compare?
- What parameters were used to define which people had ADHD? SUD? Do you think this method is reliable/sufficient?
- What assumptions about shared genetic and environmental factors do the authors make in the statistical analysis?
- What is a sensitivity analysis? What was the goal of performing this type of analysis? Why did the authors perform more than 1 sensitivity analysis?

Results

- How do you interpret the odds ratios presented in Table 1? What do they tell you about comorbidity in patients with ADHD?
- What are the major findings of the study?
- Tables 3 and 4 both present the outcomes after exclusion of individuals with different psychiatric disorders (in Table 3 individuals with bipolar disorder and schizophrenia are excluded, in Table 4 individuals with depression and conduct disorder). Why are these data presented separately?

- Do you find the data presented in Tables 2, 3, 4, 5, and 6 to be consistent with each other? What do you make of that?
- In the tables, the authors make a distinction between “substance use disorder” and “drug abuse.” How are these terms defined in the paper?

Discussion

- What do you take away from this study?
- The authors state that their data “support the hypothesis that the association between ADHD and SUD is explained by shared familial risk factors rather than harmful effects of ADHD medication.” How do they make this conclusion? Do you agree?
- What do you see as the limitations of the study?
- What are the clinical implications of this work?

Post-Guide

Skoglund, C *et al*. 2015. Attention-Deficit/Hyperactivity Disorder and Risk for Substance Use Disorders in Relatives. *Biological Psychiatry* 77:880-886.

Take Home Summary

This article describes a case control study utilizing very large Swedish national databases to examine the co-occurrence of ADHD and substance use disorders (SUD) in people with ADHD and their relatives. Specifically, the authors were interested in determining if ADHD and SUD co-occur independently of treatment of ADHD with stimulant medications. Using these databases, the authors identified 62,015 people with ADHD based on diagnosis or prescription of medication for the treatment of ADHD. The same databases of diagnoses was used to identify patients with SUD, and drug abuse was determined by diagnosis or prescription of buprenorphine or methadone. Each case was matched to 10 unaffected controls based on age, sex, and residential factors but who were not diagnosed with or treated for ADHD. Table 1 presents the data for ADHD probands and matched controls. The rates of SUD (odds ratio 10.8), drug abuse (OR 19.2), alcohol use disorder (OR 8.3), bipolar disorder (20.1), schizophrenia (OR 6.9), depression (12.8), and conduct disorder (31.4) were all substantially elevated in people with ADHD. The risks of several psychiatric illnesses were also higher in unaffected first degree relatives of people with ADHD, with odds ratio of approximately 2 for substance use disorders, drug abuse, and alcohol use disorders in the parents and siblings of those with ADHD. The OR was highest for parents (2.2), slightly lower for full siblings (1.8), and lowest for half-siblings (1.4). These findings were very similar in several different sensitivity analyses performed by the authors to assess the robustness of the findings. The findings were also very similar after exclusion of those with bipolar disorder, schizophrenia, depression, and conduct disorder. The OR for SUD, drug abuse, and alcohol use disorder were similar in maternal and paternal half-siblings of those with ADHD, which the authors interpret as evidence that biological factors outweigh environmental factors in elevating the risk of SUD among relatives of those with ADHD. The authors conclude that the association of SUD with ADHD is based on biological aspects of ADHD and not based on treatment with stimulant medication, as this risk remains elevated among relatives without ADHD and thus not treated with stimulants. This conclusion is also based on the data demonstrating that the risk is higher among closer relatives (full siblings) and lower among relatives with less common genetic material (half-siblings).

Pre-Guide

J Song et al (2017). Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *American Journal of Psychiatry* 174(8): 795-802.

Reasons for choosing this article

- This article covers a clinically relevant question: are certain medications effective for decreasing suicidal events in patients with bipolar disorder?
- This article addresses a clinically relevant question about treatment but is not a randomized controlled trial, which allows us to think about when and how we can answer questions that may not be amenable to study in an RCT.

Background

- What is your experience in using lithium and valproate in the treatment of bipolar disorder? Can you think of a time in which you've chosen one of these medications over the other in a patient with bipolar d/o? What factors did you consider in making this decision?
- How much was known about the effect of lithium on rates of suicide prior to this study? What gap in knowledge were the authors trying to fill?
- What was the study hypothesis?

Methods

- Where was this study performed? Why?
- What do the authors mean by a "within-individual" analysis? Why did they use this strategy?
- How did the authors determine when a patient was taking a medication? How did they determine who had bipolar disorder? Do you agree with how they made these determinations?
- What's the difference between "suicide" and "suicide-related events"?
- What is a sensitivity analysis? Why did the authors perform these? Specifically, what were they trying to accomplish with the analysis involving (1) thyroid medications and (2) bone fractures?

A technical point from the Background:

In the background, the authors make the following statement: "Observational studies avoid the ethical and logical problems encountered by randomized controlled trials and, additionally, have the advantage of large sample size with long-term follow-up, offering adequate numbers of rare suicide-related events. Nevertheless, observational pharmacoepidemiological studies are highly susceptible to confounding by indication, that is, patients are selected for a medication based on their risk for the outcome." What does this mean? In your own words, describe what "confounding by indication" describes and how it applies to the question studied in this manuscript.

Related questions to consider: What are the ethical problems with doing an RCT on this topic? What are logical and logistical problems the authors refer to? How is it that observational studies can have a larger number of patients and a longer follow up time? What are the disadvantages and biases associated with observational studies?

Results

- How many patients were included in the main analysis? How many of them were ever treated with lithium? With valproate?
- What was the overall number of suicide-related events? Was this number consistent with what you expected?
- What was the effect of lithium on suicide-related events? How did this compare to valproate?
- In your own words, describe the data presented in Table 1. Was the effect of lithium different in different groups of patients?

Discussion

- What do you take away from this study?
- In the last sentence, the authors state that "lithium should be considered as a suicide prevention strategy." Do you agree? How would you balance the potential lethality of a lithium overdose compared to the benefits described in this paper?
- The authors do not consider medication dose or adherence in their analysis. Does this affect your interpretation of the data?
- What do you think might be the mechanism of action on the beneficial effects of lithium on suicide? How would you test that?

Post-Guide

J Song et al (2017). Suicidal Behavior During Lithium and Valproate Treatment: A Within-Individual 8-Year Prospective Study of 50,000 Patients with Bipolar Disorder. *American Journal of Psychiatry* 174(8): 795-802.

Summary

In this interesting paper, the authors examine the old question of whether lithium treatment is truly related to a decreased risk of suicide attempts. Generations of psychiatry trainees have learned this "fact" about lithium but as with much older information about psychotropic drugs, contemporary investigators sought to validate the finding using more modern and more rigorous research methods.

To do this, they use medical record data from Sweden. Unlike the U.S., many small European countries have centralized medical records which provide opportunities for epidemiologic analysis of questions like this. The authors used computerized methods to extract data on patients who had multiple contacts with health care providers for a diagnosis of bipolar disorder. The authors were able to link data from the medical record, pharmacy records, and death records to identify medications prescribed and suicide related incidents. Data covered children as young as 15, and the investigators were able to track data on individuals for a period of just over eight years.

The investigators focused on patients who received either lithium or valproic acid, using pharmacy records to divide the study period for each subject into three month windows bracketed by drug prescriptions, in which they were classed as on neither drug, on valproate alone, or on lithium (without or with valproate). In the primary analysis subjects were compared with themselves – that is three month periods on lithium were compared to three month periods on neither drug or on valproate alone. The methods used to do this are fairly complex, and so they also used a more conventional analysis in which patients were compared on the basis of lithium use. They considered demographic variables and estimate of clinical severity and history of suicide in the analysis.

They found results that support that lithium does have an effect on suicide risk. Overall it appeared lithium use could prevent suicide incidents by about 12%. There were some interesting findings for specific groups of patients or types of events – though recall that subgroup analysis *are a priori* less reliable than the main analysis. In particular, a more rigorous definition of suicidal events requiring evidence of intent strengthened the effect of lithium, whereas the effect was unclear for events where the intent was too. Patients with bipolar I did not benefit as much as patients with bipolar II; patients with bipolar alone and with mixed episodes had unclear benefit. They found no reduction in risk associated with valproate exposure. The investigators used the incidence of bone fractures as a statistical comparator under the assumption that this rate would not be effected by Li use.

The analysis between groups of patients found similar results as the within-subject analysis. The authors acknowledge there are limitations to the data they were able to collect – in particular the inability to know whether patients were actually taking medication – but overall this was a strong validation of the use of lithium to minimize suicide risk.

Technical Point

Confounding by indication is one of the most tricky roadblocks to good epidemiology research examining the effects of treatment. In essence the problem is that doctors don't prescribe randomly, our prescription habits are biased. Of course, mostly this is a good thing – we are biased to prescribe antidepressants for depression, for example – but it causes problems in analyses like these. It is also, unfortunately, exploited by people performing these analyses to promote an agenda. A good recent example is an article published showing that antidepressants increase mortality. If you follow this literature you would know these authors have a history of publishing "anti" psychotropic medication articles, which often take advantage of poor understanding of confounding by indication to make a splash in the media. In the case of the antidepressant publication, confounding by indication means that no patients are included in the study who are exposed to antidepressants who aren't also effected by depression (its questionable whether there are such patients at all in any numbers, though you might be able to find some with anxiety alone perhaps). In other words all the finding could be attributed to depression rather than antidepressant exposure. You might object that patients with depression who aren't medicated would also make a good control group but here's where the confounding becomes even more important. In the real world, patients with less serious depression may not be medicated, but very sick patients – those with more comorbidities, hospitalizations, suicide attempts – will almost always be. In other words, we don't prescribe antidepressants randomly! There's already a lot of evidence that depression leads to poor health outcomes and one of the few clinical findings shown again and again is that the sickest patients do the worst over time. It's easy to miss this if you aren't thinking critically about the design of a study and how the analysis was conducted, and easy to turn a blind eye to it if you already harbor stigma against psychotropic medication.

Fortunately in this study, if doctors were prescribing lithium to patients expressing suicidal thoughts in the belief that this would help them, that would be expected to decrease any apparent effect on suicidal behavior lithium might have, so its unlikely to have biased the findings in the direction the authors expected.

Pre-Guide

The TADS Team. (2007) The Treatment for Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psychiatry* 64(10): 1132-1144.

Reasons for choosing this article

- The article provides important data on the treatment of adolescent depression with CBT, fluoxetine, or the combination, a common clinical issue.
- This article also provides important data on the relative safety of each of the treatment modalities.
- The design of the study allows for a discussion of the benefits and limitations of "effectiveness" trials, which are an important component of the evidence-base for a variety of treatments used in several psychiatric illnesses.

Background

- Which aspects of the TADS trial does this paper present? What data from TADS had been published prior to this article?
- What is the authors' hypothesis at the outset of the study? How does that hypothesis apply to the current report?

Methods

- Who were the study participants?
- The authors notes that all participants and at least 1 parent provided "informed consent/assent." What is the difference between consent and assent in this context? Which term applies to the parents and which to the participants?
- How were the treatments given to the patients (how was fluoxetine dosed, what was the number/frequency of CBT sessions, etc.)? How does this compare to how these treatments are delivered in typical outpatient practice?

A technical point from the methods:

- The authors describe different types of analyses for the “scalar outcome measures” vs the “binary outcomes.” In this study, which measures are scalar outcomes measures and which are binary outcomes? What are the differences between these types of data? How are they analyzed differently?

Results

- What are the major findings of the study with regards to the efficacy of fluoxetine alone, CBT alone, and combination therapy? How do the efficacy findings change over time? (hint: Figure 2)
- The authors describe the participants in the sample as having "moderate to moderately severe" depression. How did they make this determination? Do you agree with this characterization?
- How did participant drop-out compare across the treatment groups over the 36 week period? What do you make of any observed differences?
- How do you interpret the relative effective sizes and number-needed-to-treat (NNT) data presented in Table 3?
- How common was suicidal ideation at baseline? Overall, how did suicidal ideation change over time during the 36 weeks of the study?
- How was "suicidal event" defined? How do the data for suicidal ideation and suicidal events compare?

Discussion

- What do you take away from this study?
- The authors argue that the TADS data are widely generalizable. Do you agree?
- Suppose you are seeing a 16 yo patient with moderate MDD. In discussing treatment options, the patient's parents say they do not want their child treated with fluoxetine because they have heard that "Prozac causes kids to become suicidal." Given the TADS data, how would you respond to their concern?
- The authors describe TADS as an "effectiveness trial" and contrast it against "comparative treatment trials." What do these terms mean, and how do the goals of each type of trial differ? Do you agree with the authors that TADS is an effectiveness trial?
- The authors state: "we readily acknowledge that it is impossible to conclude that patients would not have reached equivalent week 36 outcomes simply because of the passage of time without a placebo group or, better, an untreated control group, both of which were considered unfeasible for ethical and practical reasons." What do you make of the lack of a placebo or untreated group? What would have been added to the trial with the inclusion of such a group? Why might including such a group have been unfeasible or unethical?

Post-Guide

The TADS Team. (2007) The Treatment for Adolescents with Depression Study (TADS): Long-term Effectiveness and Safety Outcomes. *Arch Gen Psychiatry* 64(10): 1132-1144.

Take Home Summary

The Treatment for Adolescents with Depression Study (TADS) is a large effectiveness trial comparing fluoxetine alone, CBT alone, or the combination of fluoxetine and CBT in the treatment of adolescents with major depression. This study was done in the heyday of large effectiveness trials, and following the addition of a black box warning to the FDA required labeling for antidepressants in teens and young adults. Many aspects of the design were influenced by the need to closely examine the efficacy of antidepressants in this age group given the possibility of increasing suicidal behavior. The primary TADS end point was at 12 weeks, and the primary outcome paper showed that combination therapy was superior at that time point. The present report describes the long-term (36 week follow-up) outcomes of treatment efficacy and safety with a focus on suicidal ideation and behavior. The study enrolled 327 patients aged 12-17 with MDD and tested the hypothesis that combination therapy would show greater benefit faster than either treatment alone and that the advantage of combination therapy would be maintained throughout the study.

Subjects were recruited from 13 sites around the US and were randomized after initial assessment of eligibility. Blinded fluoxetine (or placebo) was started at 10 mg and was incrementally increased to as much as 60 mg per day during the first 12 weeks of the trial; the dose was decreased if a subject experienced intolerable side effects. CBT was provided in 15 1-hr sessions over the first 12 weeks, with gradually decreasing frequency over the remainder of the trial. Primary outcomes were determined by scores on the Children's Depression Rating Scale-Revised (CDRS-R) and the Clinical Global Impression-Improvement (CGI-I) scale. Suicidal ideation was monitored using the Suicidal Ideation Questionnaire-Junior High School Version (SIQ-Jr) and suicidal events were determined by the Columbia rating scale. Study participants were average age 14.6 years and had experienced depression for an average of 75 weeks; at baseline 28% had at least minimal suicidal ideation.

For this paper of 36-week outcomes, data from 243 of the 327 participants were available, due to the exclusion of subjects on placebo who were unblinded at the end of 12 weeks, as well as drop out. Medication treatment was thus unblinded for the long-term phase of TADS. In all treatment groups, the majority of patients showed improvement in symptoms of depression, with 86% of participants in the combination group, 81% in the fluoxetine group, and 81% in the CBT group demonstrating a response to treatment at 36 weeks, rates that were not statistically different between groups. Overall, suicidal ideation decreased in all groups over the study; however, the rate of suicidal ideation decreased more in the CBT and combination groups than in the fluoxetine alone group. Additionally, more suicidal events occurred in the fluoxetine alone group than the other groups. From these data, the authors conclude the combination therapy (CBT+fluoxetine) is the most effective treatment for adolescent depression. While medication seemed to produce faster response than CBT, CBT seemed to offer a buffer to the possible increase in suicidal behavior seen with fluoxetine, at least in the long term. Most importantly, the study established that antidepressants could be used effectively in this age range, while validating the existence of the controversial increase in suicidal behavior, the benefit exceeded the risk.

Technical Point:

There are two major ways to analyze outcomes in clinical trials: discrete and continuous. In this paper they use the terms 'binary' and 'scalar' respectively but these refer to the same thing – whether we divide subjects into groups or look a continuous distribution of score (or change in scores) on an outcome measure. In mental health there is ambiguity about which of these to choose. In some disciplines and settings its 'obvious' – you either have cancer or you don't, you are alive or dead, or there is a clearly continuous measure such as blood pressure readings for analysis. In mental health its rarely straightforward – depression varies in severity within individuals over time and may be subthreshold in terms of diagnosis but still clinically significant. Perhaps more importantly, there are no obvious units in which to measure symptoms (unlike with blood pressure). Researchers have therefore coalesced around standard practices. The most common discrete or binary measures are "response" and "remission." Response is typically defined as a 50% reduction in symptoms and is not measure specific. Remission is defined as a score below a measure specific threshold, e.g., "a score of less than 10 on the Hamilton Depression Rating Scale." It's important to understand that remission does not mean symptom free and that it depends on the measure being used in the study. Imagine a hypothetical subject who enters the study with a score of 16 on the Hamilton. If at the end of the study they have an 9, they would be classed as a remitter but not a responder, which may seem, and is, counterintuitive. For this reason usually only patients are entered into the study for whom their baseline depression severity is at least twice the remission cut-off. Continuous outcomes typically look at the score itself or the difference between the initial and final scores.

Very different statistical methods are used to evaluate these types of outcomes. Typically, a chi-square test is used to assess if the number of subjects with a binary outcome is different between groups. T-tests, ANOVAs, and linear models are appropriate for continuous data– these can be correlation, regression, or more complex models such as the Generalized Estimating Equations used in this paper. We will cover these methods in other articles, here we want to emphasize that you understand that researchers must design studies to work best with one of these measure types and use statistical methods to match. Most large trials report both types of outcomes but one should clearly be primary and study design elements such as the sample size should be determined by the main methods used.

Pre-Guide

M. J. Telch *et al*. (2014). Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *American Journal of Psychiatry* 171:1091-1098.

Reasons for choosing this article

- This article presents a different type of trial—a human laboratory study in a non-clinical population.
- This article allows us to discuss different types of learning and the implications of different types of learning for the treatment of anxiety disorders.
- This article puts a somewhat different spin on the idea of combining medication and therapy for the treatment of anxiety disorders.

Background

- The article discusses several different types of learning—extinction, contextual memory, and consolidation. What is the definition of each term?
- What is the rationale for using methylene blue, both in terms of safety and proposed mechanism?
- What is the study hypothesis? Why do the authors propose different effects of methylene blue based on the efficacy of the extinction training?

Methods

- Who were the study participants? Why do you think the authors recruited a non-clinical sample? What are the benefits and limitations of the study population?
- What was the extinction training paradigm? When was methylene blue administered? When were outcome measures collected in relation to the training?
- How was the efficacy of extinction training measured? What did the authors do to assess for a non-specific enhancement in memory?

A technical point from the methods:

- On page 1092, the authors state: “the 260-mg methylene blue dose corresponds to the 4 mg/kg dose shown to be effective in previously published preclinical studies.” In what ways is this statement problematic?

Results

- The first sentence of the results state: “the mean fear level was 73.0 at the first exposure and 23.5 at the last exposure.” What do these number tell you about the efficacy of the extinction training?
- Describe the results as presented in Figure 1. How would you describe the relationship between methylene blue treatment and fear level at 1 month follow-up?
- How did the authors divide participants into low, average, and high end fear? Why did they perform this stratification? Does their method for stratifying seem reasonable to you?
- The authors asked two questions, one about fear memory and one about contextual memory. How do these questions differ? How do they relate to each other? (i.e., what’s the difference between Figure 1 and Figure 2, and why are both sets of data included?)

A technical point from the results:

- The authors note that there were no serious adverse events. What is the standard definition of a serious adverse event?

Discussion

- What do you take away from this study?
- What do the authors mean when they describe methylene blue as a “cognitive enhancer?”
- Why do you think the editors considered these findings worthy of publication in *AJP*? Do you agree?
- Do you think there will be a role for methylene blue treatment in clinical practice? What might that look like?

Post-Guide

M. J. Telch *et al*. (2014). Effects of Post-Session Administration of Methylene Blue on Fear Extinction and Contextual Memory in Adults with Claustrophobia. *American Journal of Psychiatry* 171:1091-1098.

Take Home Summary

This article describes a human laboratory study of the efficacy of methylene blue treatment to enhance extinction of claustrophobia (fear). Extinction of fear describes the observation that people become less afraid of things when they are repeatedly exposed to them—in the case of this study, people with sub-clinical claustrophobia remained within a small, enclosed space for six 5-minutes sessions. Participants were given either methylene blue or a placebo to take 3 times in the 24 hours after the training. Level of fear in an enclosed space was measured one month later. The authors hypothesized that the effect of methylene blue treatment would depend on the level of fear immediately after the extinction training—those who had low fear at the end of the training and methylene blue would show an enhancement of this effect at one month (compared to those who had low fear and received placebo), while those who had high fear at the end of the training and received methylene blue would show even higher levels of fear one month later. Via its activity in the mitochondria to enhance cellular energy production, methylene blue is thought to enhance the consolidation of the information that has just been learned. In this way, methylene blue would act as a cognitive enhancer. The study results matched what the authors predicted. Importantly, the authors also demonstrated that the effects of methylene blue on fear were separate from effects on contextual memory—memory of the environment in which the extinction training was performed. During the extinction sessions, numbers were displayed in the enclosed space, but patients were not instructed to remember them. At the follow-up, patients were asked to remember those numbers. The patients who received methylene blue demonstrated better memory for the numbers than those given placebo, regardless of the fear level.

The authors propose that methylene blue may be a clinically useful agent for augmenting exposure therapy for the treatment of specific phobias, although they caution that this medication should only be given to patients in whom the exposure therapy has led to decreased fear levels. The conclusions of the study are limited somewhat by the study population—a non-clinical sample of mostly female undergraduate. It is now known if these results would generalize to a clinical sample with claustrophobia or other specific fears.

Regarding the technical points from the pre-journal club guide:

1. Briefly, people are not rats. Because of the differences in body surface area, volume of distribution, liver metabolism, etc., there is not direct correlation between an effective dose in preclinical rodent studies and human studies.
2. A serious adverse event (SAE) is generally defined as anything requiring emergent medical treatment, regardless of whether or not the event was related to the study. Per FDA guidelines,

SAEs include death, life-threatening conditions, hospitalizations, disability or permanent damage, and congenital anomaly/birth defect.

Pre-Guide

MH Trivedi *et al*. (2006) Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 354: 1243-1252.

Reasons for choosing this article

- This paper is one of the major outcomes papers of the STAR-D trial, which is the major effectiveness trial for the treatment of depression. There is no single primary outcomes paper from STAR-D. Instead, the authors published this article about the effects of the augmentation strategy and a separate, accompanying paper about the patients who were switched to a different medication after lack of response to citalopram.
  - For an overview of all of the phases of STAR-D, see: Warden D. *et al*. 2007. The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.
  - The medication switch paper is Rush et al. (2006) Bupropion-SR, Sertraline, or Venlafaxine-XR after Failure of SSRIs for Depression. *N Engl J Med* 354:1231-1242.

Background

- At the time of the study, what was known about treatment of major depression in patients who did not respond to the first medication trial?
- What were the goals of STAR-D? What was the goal of this part of STAR-D?
- What do you think was the study hypothesis?

Methods

- Who were the study participants? How did they qualify for this part of STAR-D?
- What medications at what doses were studied? What do you think of the choice of these medications?
- How was the efficacy of the medication measured? What is meant by the terms response and remission?
- How long did treatment in this phase of STAR-D last?

A technical point:

- The authors use two different measures of depressive symptoms: the HAM-D, which is a clinician-rated assessment, and the QIDS-SR, which is a patient self-report. What are the differences between clinician-rated and self-report measures? What are the advantages and disadvantages of each type of assessment?

Results

- How many patients were randomized in this phase of STAR-D? Do you consider this to be a large study? When the authors say that "most of the patients who were randomly assigned. . . .had accepted their assignment to receive only the two augmentation medications," what feature of STAR-D are they talking about?
- How would you describe the baseline severity of illness of patients in this study? (Table 1)
- The authors note that the "medications were administered in effective doses and were provided for adequate durations of time to detect benefit." Why do they make of a point of saying this? Do you agree with this statement?
- What was the effect of each medication on the primary outcome? On what outcome measures were the results the same for buspirone and bupropion, and on what outcomes did the treatment effects differ?
- The authors present the outcomes in terms of remission, response, percent change in score from baseline, and final score on symptom severity rating scales. Which of these outcomes do you think is the most meaningful? What is the value in presenting all of these outcomes?

Discussion

- What do you take away from this study?
- What augmentation strategies have you used clinically?
- On page 1251, the authors discuss the fact that this is not a placebo-controlled study. Is this study weakened by lack of a placebo? What are the reasons not to use a placebo?
- The final sentence of the paper is "These results raise the question of whether to use augmentation agents (or other treatments in combinations) as first-line treatment in an attempt to achieve greater remission rates sooner in more patients that with SSRIs alone." What data presented here support the idea of starting with an augmentation/combination medication strategy? Should we routinely start patients with major depression on a medication combination?

Post-Guide

MH Trivedi *et al*. (2006) Medication Augmentation after the Failure of SSRIs for Depression. *N Engl J Med* 354: 1243-1252.

Take Home Summary

STAR*D has become so ingrained in psychiatry that it can be hard to appreciate in retrospect how major it was at the time. Augmentation – adding a medication to an antidepressant given an adequate trial with only partial response – is a very common strategy for treating patients who have demonstrated treatment resistance. (It should be contrasted with combination treatment, in which a second medication is started early, before the full efficacy of the first can be determined.) Augmentation is typically accomplished by starting an SSRI (or perhaps an SNRI) and then adding a second medication has a different mechanism of action. However, until STAR*D, psychiatrists were flying blind from an evidence perspective; no one knew with certainty that we were improving outcomes for patients with depression by adding medications or by how much we were increasing the risk of side effects. This question was addressed in the second step of STAR*D and reported here.

In the first STAR*D step, adults with depression were treated with citalopram for 12 weeks, with anyone who achieved remission able to enter the follow-up phase of the study. Those who did not remit by 12 weeks were eligible for step 2 which had a number of options: switch to a different antidepressant, receive augmentation, or add psychotherapy. Subjects were able to choose which of these options they were willing to be randomized to; most chose only a subset of the options. 565 patients ended up being assigned to the augmentation group, in a process called pseudo-randomization since subject preference was accounted for. Once assigned to augmentation, they were randomized to either bupropion SR BID or buspirone BID with flexible dosing. Like all of the STAR*D treatments, they received medication openly (no blinding). For the primary outcome, depression severity was assessed over the phone, by centralized raters who did not know what treatment group the subjects were in, so the assessments were blinded.

The dosing of medications in this type of study is often a point of criticism, and it was here as well. The mean dose of total daily bupropion was 267mg and the dose of buspirone 40.9. For most clinicians, 150-300mg of bupropion seems typical, but the buspirone dose seems low for a total daily dose. More concerning is that buspirone is meant for TID dosing – most commonly, if you prescribe it at all you would start at 10mg TID, going up to 15mg or 20mg TID. The dosing here suggest most subjects were getting *15mg or 20mg BID*. Although there were a number of studies suggesting some patients do just as well on BID dosing, mostly done in the 1990s, there is still a possibility that the choice of dosing impacted the outcome. In particular if total dose was limited by splitting the dose into only two pills, then buspirone may have looked unfavorable compared to bupropion.

The results based on the a priori primary outcome at 12 weeks, remission on the Hamilton Rating Scale, showed no difference between the groups. However, subjects also provided symptom ratings on the Quick Inventory of Depression Symptomatology (QIDS-SR) and these results showed a pattern that favored bupropion. Both the change in depression score and the final depression score on the QIDS significantly favored bupropion, though we should note that neither of these p values would likely survive correction for multiple testing. The only large difference is that about twice as many people stopped buspirone due to intolerance compared to bupropion.

Overall, the major take-home of this paper is that augmentation is a successful strategy for treating depression in people who have failed to remit after 12 weeks on a single agent. As far as choice of agent, bupropion remains far more common than buspirone in practice today, possibly because of the perception that it has higher side effects.

Technical Point

A major hurdle to clinical psychiatric research is measuring outcomes. Unlike many areas of medicine, psychiatric symptoms are not obviously quantifiable, and they are largely subjective. Thus, there is an art to measuring symptoms which has become more important as it has moved into the clinical realm. It turns out to be very complex, as slight differences in the way questions are asked can significantly affect the answers. One of the biggest debates in clinical trials is whether we should have subjects complete forms on their own – self reports – or be asked questions by a trained rater in order to quantify symptoms. In both cases there is a risk of bias. Patients may have 'error' in their self-assessments, in that they may not even view themselves in the frame of an assessment. For example, subjects who are clearly very depressed circle "I do not feel sad" on the QIDS-SR because they described their mood as "low. " Subjects may feel that the answer choices don't reflect their symptoms, they may attempt to edit the question stems, try to rate "2.5," or otherwise change answer choices. Even on computerized assessments, subjects may provide consistently high or consistently low scores, depending on their self-perception and goals.

For these reasons many experts in psychometrics feel that clinician administered assessments may be better than self-rated. Clinicians may be better able to determine symptoms severity (particularly in relationship to severity other patients) and be better able to translate the patient's report of symptoms onto the scale measures. However, raters also may suffer from various biases. In trials, raters may be biased towards the success of the trial and follow the expected pattern of high depression scores at enrollment and low scores at exit. Several studies comparing clinician to self-rated scales have found that at study baseline, raters tend to give higher depression severity scores to subjects than subjects give to themselves. For this reason, large efficacy studies like STAR*D often used blinded raters who did not know where a subject was in the study.

This particular paper is a good if murky example of why there is a debate over self- and clinician-rated assessments. It's hard to say for sure why the results of the QIDS analysis showed more of a difference between the two treatments. One explanation is the real differences between the scales. A big one is that the QIDS was meant to be more sensitive to change in patients with relatively mild depression to begin with, while the Hamilton was designed for very sick inpatients. Because the team behind STAR*D developed the QIDS scale, there was criticism that perhaps those results were played up because they advertised the QIDS, or perhaps they were just 'p hacking' – assessing the data different ways until something significant is found. Perhaps subjects noticed something about themselves that raters didn't? In any case, since this time most studies collect both self and clinician rated assessments! So at least we go on comparing them.

Pre-Guide

Uher R *et al*. 2009. Genetic Predictors of Response to Antidepressants in the GENDEP Project. *The Pharmacogenetics Journal* 9:225-233.

Reasons for choosing this article

- This article reports the primary outcomes of GENDEP, a large, carefully done study that aimed to identify genetic variations associated with response to antidepressant medications. This article lets us review several important concepts for understanding human genetic studies.
- This article can help us identify the potential benefits and limitations of performing tests for genetic variations and making medication decisions based on those test results.

Background

- How many authors are on this paper? Does that number seem high? Why do you think that is?
- What type of pharmacogenetics work had been done prior to this study? What gaps in knowledge were the authors hoping to fill?
- Remind yourself of the definitions of the following terms: intron, exon, coding region, single nucleotide polymorphism (SNP)
- What genes were included on the list of tested candidates? Does this list seem appropriate?
- What were the authors' hypotheses?

Note: This journal organizes manuscripts such that the methods section is presented at the end. You may read the article as written, or you may read the methods first and then the results.

Methods

- What do the authors mean by “part-randomized”? What are the implications of structuring the randomization scheme in this way?
- Why did the authors pick the two medications they did? Why did they study only two medications? Why didn’t they include a placebo? What do you think about these decisions?
- Why were participants not allowed to take any other psychotropic medications?
- How many comparisons were performed?
- How big was the final study population?

A technical point from the Methods:

The methods used in genetic studies are quite specific to the field. In what ways are different genes not independent from each other, and what statistical issues are associated with this? Consider the following terms: linkage disequilibrium, haplotype, conserved, polymorphic, HapMap, Tag SNP, false discovery rate.

Results

- Which genes were associated with overall medication response? Escitalopram response? Nortriptyline response?
- How do you interpret Figure 1? I.e., what's on the x-axis, what's on the y-axis, what do the dots mean, etc.?
- Define each of the column headings in Table 1. Which of the listed findings are statistically significant?
- What do the authors mean by "nominally significant" results?
- What does it mean for a gene difference to account for 1% of the variance in response (the authors talk about this a bit more in the discussion)?
- What did the authors find when they tested previously reported gene SNPs for effects?

Discussion

- What do you take away from this study?
- The authors state: "the distribution of multiple positive findings supported the hypothesis that pharmacogenetics associations are specific to antidepressant mode of action." Why do they make that conclusion? Do you agree?
- How do the authors explain difference they found from previous studies? Is their argument reasonable?
- How do understand the difference between statistical significance and clinically meaningful difference in this type of study? Based on the results presented here, should we be testing patients for SNPs and making medication decisions based on the results?
- On page 229, the authors state: "clinically useful pharmacogenetics prediction could be achieved either by finding genetic markers with much stronger effects or by combining a number of weakly predictive markers." Which do you think is more likely?

Post-Guide

Uher R *et al*. 2009. Genetic Predictors of Response to Antidepressants in the GENDEP Project. *The Pharmacogenetics Journal* 9:225-233.

Article Summary

GENDEP is a large pharmacogenetic study of candidate genes and antidepressant outcome. The study recruited over 800 subjects with MDD, and pseudorandomly assigned them to either escitalopram or nortriptyline treatment for 12 weeks. The inclusion and exclusion criteria were typical for clinical depression trials, except that ethnicity was limited to white Europeans, to provide a more genetically uniform sample.

Single Nucleotide Polymorphisms (SNPs) in 10 genes in three categories--1) serotonin system 2) norepinephrine system 3)"final common pathway"--were examined. The authors predicted that variants in the first two categories would predict outcome with escitalopram and nortripyline, respectively, and that variants in the 3rd category might effect outcome from both drugs.

Serotonin		Norepinephrine		Common Pathway	
HTR1A	Receptor gene	*SLC6A2*	Transporter gene	*NR3C1*	Cortisol receptor gene
HTR2A	Receptor gene	ADRA2A	Receptor gene	*FKBP5*	gene for cortisol receptor 'helper'
TPH1	Enzyme gene			*BDNF*	Neurotrophin gene
TPH2	Enzyme gene			*NTRH2*	BDNF receptor gene

Results of the analysis were overall consistent with the hypotheses of the study. Three variants in the gene for the serotonin 2A receptor, *HTR2A*, were associated with escitalopram over nortriptyline, and two variants in the gene for the norepinephrine transporter, *SLC6A2*, were associated with response to nortriptyline over escitalopram. Finally three variants in the cortisol receptor gene *NR3C1* were associated with outcome regardless of treatment. However, when correcting for the number of hypothesis specific tests, only two of the *HTR2A* SNPs had a significant association with outcome. Each of these variants, though significant, only accounted for about 1% of the difference in outcome between the groups.

The authors also divided assessments into subscales and compared outcomes using these to genetic variation. This was done because it is thought that some symptoms might be more genetically determined than others. These results were very similar to the overall results for the "core mood" subscale, with one additional association for a variant in BDNF reaching significance for cognitive symptoms. They also looked specifically at replicating findings from older studies, but were unable to do so, even though in some cases other variants in the same gene were significant.

Comments

The most critical aspect of GENDEP is that is was a **hypothesis driven** study. It was designed based on prior literature, almost all of which was post hoc analysis of clinical trials such as STAR*D that were not designed to assess genetic outcomes. Most of the prior literature did not have comparison groups for outcomes and therefore made it impossible to truly assess drug specific effects as opposed to general factors favoring a good outcome.

Technical Point

Genetics studies use a lot of specialized methods and terminology. Most of these arise from the fact that DNA variants are not inherited individually – or from a statistical perspective, independently. It's very important to understand that genetic variants never act alone. This is the primary reason that studies examining the effect of genes on outcome attempt to use closely related subjects.

Essentially, blocks of DNA (haplotypes) are inherited together, so the "local" environment of a SNP may be very different in different populations. For example, it has been found that some SNPs have opposite effects in Han Chinese and Caucasian samples. If, in one population, the serotonin transporter is more efficient, serotonin reuptake may be less effected by a mutation that lowers its cell-surface expression level. This would make the effect of the mutation correspondingly harder to detect. The take home message is that **the more similar a population is overall, the more detectable differences caused by a single genetic variant will be**. This also means that in ethnically and genetically diverse populations such as in the US, genetic findings from more uniform samples may not apply.

Linkage Disequilibrium a measurement of how often two genetic variants are inherited together. If variants are in "equilibrium" they are inherited together 50% of the time (usually this means that they are on different chromosomes). DNA sections and variants on the same chromosome are almost always, therefore, in linkage disequilibrium. A value of 1 means the variants are always inherited together.

Haplotype – A (typically small) section of DNA that tends to be inherited together, that is, is in high linkage disequilibrium. This is a part of a chromosome which typically is not split up in recombination events. Variants in haplotypes are therefore mutations on a shared genetic background.

Conserved – There is not much genetic variation in DNA regions that are conserved. This implies that mutations in these regions are typically harmful and are selected out by evolution.

Polymorphic – The opposite of conserved. Polymorphic regions have a lot of variation across a population.

HapMap – a large genetic database used as a reference for tracking genetic variation in human beings, determining the values for linkage disequilibrium, and choosing Tag SNPs.

Tag SNP – a SNP chosen to represent a haplotype; different algorithms acting on different genetic databases may choose different SNPs to represent the same section of DNA. Tag SNPs may or may not have any functional significance, instead they represent a way of simplifying the analysis.

False Discovery Rate – a method used to control for multiple testing in situations in which the number of possible associations is large compared to the number of subjects. Most genetic studies use FDR correction, as the number of genes (and definitely the number of variants) is higher than the number of subjects. Here, the authors give a traditional correction because the number of tests, though large, was still smaller than the number of subjects, but also include FDR correction, which led to no significant results (Table 1).

Pre-Guide

The UK ECT Review Group. 2003. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 361:799-808.

Reasons for choosing this article

- Although now old, this article provides critical evidence on the efficacy of ECT. Given how busy our ECT and hospital services are, there is often little time to review this important evidence while on service.
- This paper lets us think about the unique features of meta-analysis.

Background

- Based on your clinical experiences, what do you know about the efficacy of ECT?
- What do you think was the authors' hypothesis?

Methods

- Define the following terms: randomized, unconfounded, controlled
- Where did the data used in the study come from?
- Which groups did the authors compare?
- What factors were used to assess the quality of RCTs? Are these reasonable criteria? Why is this important?
- The authors state that they wanted to “avoid risk of multiple testing or data-driven analyses.” What do they mean by this?

Technical Point:

- What is a meta-analysis? What are the advantages of this type of analysis? Pitfalls? What is a role of a funnel plot?

Results

- How do you interpret each Figure? What information is presented on the x-axis and y-axis? What is presented in each row? What are the diamonds? What are the bars?
- The authors report that, on average, patients who received ECT treatment vs sham ECT demonstrated a decrease in Hamilton Depression Rating Scale (HDRS) score of 9.7. Is this a big change? What is the clinical significance of this change?
- What different aspects of ECT do the authors examine? Are the data from the different comparisons consistent with each other?
- What are the findings regarding the safety of ECT?

Discussion

- What do you take away from this study?
- How robust is the evidence base supporting ECT?
- Based on this meta-analysis, is there evidence to support the statement that ECT is a life-saving treatment?
- The authors note that one limitation is that clinical trials (at the time of publication) had not investigated the common clinical practice of short term ECT followed by medication treatment to address residual symptoms and relapse prevention. How would you design a study to assess this?

Post-Guide

The UK ECT Review Group. 2003. Efficacy and Safety of Electroconvulsive Therapy in Depressive Disorders: A Systematic Review and Meta-Analysis. *Lancet* 361:799-808.

Take Home Summary

Electroconvulsive therapy (ECT) is one of the oldest, most effective, and most controversial treatments in all of psychiatry. It can it a tough sell for many patients, even those with treatment resistant depression who have not responded to several trials of pharmacotherapy. Therefore, it's important for psychiatrists to have facts on hand about the efficacy of ECT and its side effects. In addition to supplying that information, this paper is a good example of a meta-analysis because it combines data from older studies which were under-powered, and which may not have used standardized clinical outcomes, illustrating the kinds of decisions investigators need to make to produce synthesis of data.

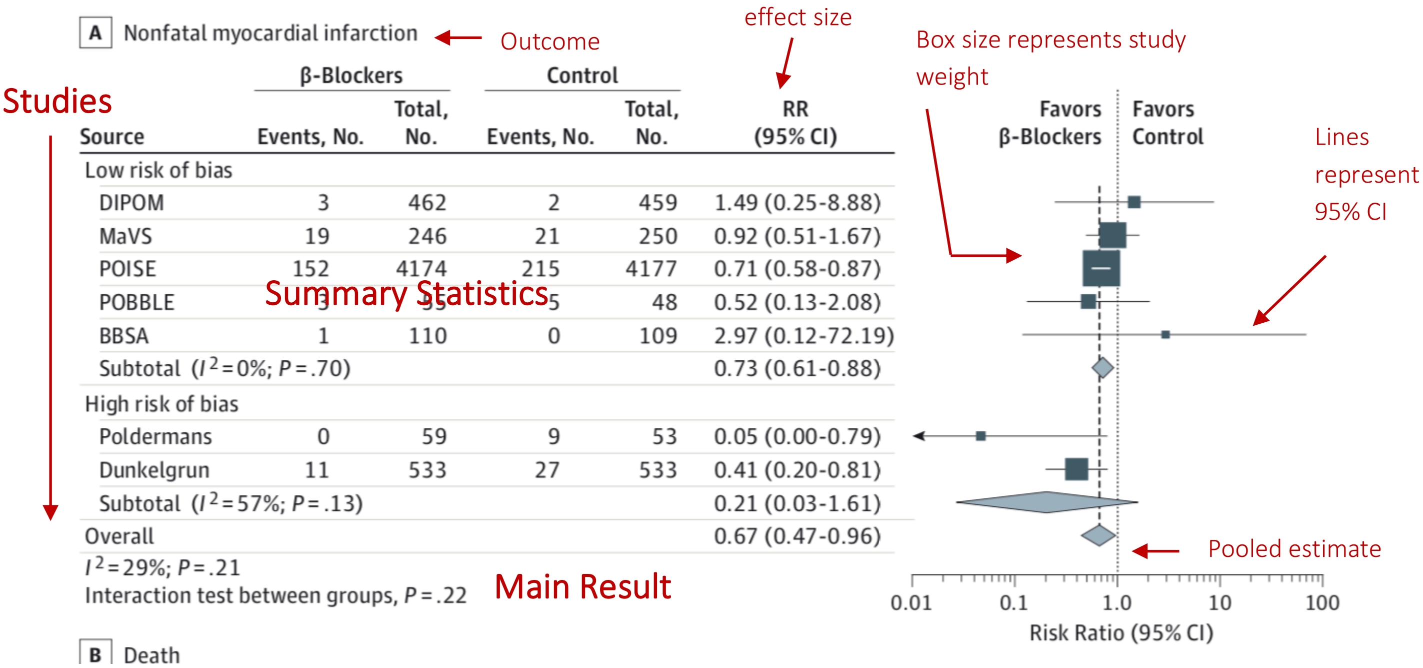
This meta-analysis is of high quality because the review group followed a rigorous procedure to find, assess, and include/exclude studies. They also used a clear set of criteria for assessing study quality, which is particularly important for these data because blinding or masking adequately can be more difficult for procedures than in pharmaceutical trials. First, assessment of outcome must be done by researchers who are not part of the treating team. Next, sham ECT treatment must be of high quality, ideally the entire process of ECT, including prepping the subject, sedating them, hooking up electrodes, waking them up, etc. is conducted *except* that there is no current delivered. This ensures factors such as the amount of time spent sedated are the same for both groups. Because drop-out may be high and bias outcomes, they also examine early study as an outcome, for additional control of bias.

The important findings were that ECT is superior to (1) no treatment (sham) with a huge effect size of almost 10 points on the Hamilton Depression Rating Scale and (2) pharmacotherapy with a large effect size of 5 points. These effects are better than most medication studies where effects may only be 2-3 points more decrease compared to placebo. They also found that higher dose but lower frequency treatment were superior (dose being the amount of current used, frequency number of treatments/week), and that sinewave stimulation and bilateral treatment were superior (but bilateral in particular was associated with worse cognitive outcomes). An important caveat to these findings is that it still may be difficult to use them to guide practice since the individual studies were difficult to compare, which is a weakness of meta-analyses in general. Examining 'high' and 'low' dose for example, there was so little consistency among the studies in how doses were defined that it would not be easy to decide on exactly what dose to use based on these results. Lastly, evidence that the ECT treatment effect was durable was modest compared to evidence of its acute efficacy.

The results are helpful in being able to confidently assert that ECT is an effective treatment for TRD (perhaps most effective treatment for TRD, though it was not compared to modern, state-of the art pharmacotherapy in most of the studies). Although side effects are high, most of the evidence supports that most patients get more benefit than harm from the treatment.

Technical point

Meta-analysis works in one of two ways. Sometimes investigators are able to obtain the original data from studies and, by choosing studies with similar methods, are able to analyze the data as one large set. Often however this is not possible, because differences in study conduct make it too difficult. In that case, methods of standardizing the outcomes are used. Usually these methods rely only on values (summary statistics) available in a published paper, like the mean, standard deviation, odds-ratio and so on. Then, the effect sizes for each study are combined to come up with a final effect size representing all the studies.



A Nonfatal myocardial infarction

	β-Blockers		Control		
Source	Events, No.	Total, No.	Events, No.	Total, No.	RR (95% CI)
Low risk of bias					
DIPOM	3	462	2	459	1.49 (0.25-8.88)
MaVS	19	246	21	250	0.92 (0.51-1.67)
POISE	152	4174	215	4177	0.71 (0.58-0.87)
POBBLE	3	55	5	48	0.52 (0.13-2.08)
BBSA	1	110	0	109	2.97 (0.12-72.19)
Subtotal (I^2=0%; P=.70)					0.73 (0.61-0.88)
High risk of bias					
Poldermans	0	59	9	53	0.05 (0.00-0.79)
Dunkelgrun	11	533	27	533	0.41 (0.20-0.81)
Subtotal (I^2=57%; P=.13)					0.21 (0.03-1.61)
Overall					0.67 (0.47-0.96)

I^2=29%; P=.21
Interaction test between groups, P=.22

B Death

Outcome
effect size
Box size represents study weight
Studies
Summary Statistics
Lines represent 95% CI
Pooled estimate
Main Result

This meta-analysis used the Standardized Mean Difference (SMD) as the measure of effect size or difference between arms in each study. The SMD is calculated by dividing the difference in mean between the arms by the standard deviation for the whole sample. Meta-analyses use forest plots to express their findings. Although you must read the methods and results sections to evaluate the quality of a meta-analysis, you can easily and quickly understand the results by looking only at the plots. The figure above is from Murad et al, JAMA (2014; full reference below), a great guide to reading and applying meta-analyses, is used here as an example. Forest plots list studies, with individual summary statistics and a plot, with bilateral symmetry, one side representing each arm. Boxes or diamonds show the effect estimates for each study. Bars or lines centered on the effect estimates reflect the 95% confidence intervals (CIs) for the studies. In a meta-analysis for which studies are weighted, the size of the box/diamond represents the weight of the study. Weight is based on the precision of the estimate, that is the inverse of the CI. In our ECT meta-analysis as well as this example figure, larger studies typically have more precise CIs and, therefore, greater weight. At the bottom the pooled effect size will be shown as a diamond whose width represents the pooled CI. If the diamond does not touch the center line, then a significant pooled effect has been found, favoring the treatment arm for that side.

For more guidance on reading a meta-analysis, see: MH Murad *et al*. How to Read a Systematic Review and Meta-Analysis and Apply the Results to Patient Care: Users' Guides to the Medical Literature. *JAMA* 2014; 312(2):171-179.

Pre-Guide (PGY2-4)

Warden D. *et al*. 2007. The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.

Reasons for choosing this article

- STAR*D, the Sequenced Treatment Alternatives to Relieve Depression trial, is a major effectiveness trial of antidepressants. The trial had many steps and parts, and this review article provides a summary of STAR*D outcomes.
- By reading this review article, we can look at STAR*D comprehensively and consider how well the trial met its goals.

Background:

- What were the goals of STAR*D? What clinical scenarios did STAR*D attempt to clarify?
- The S of STAR*D stands for "sequenced." What does this describe?

Methods:

- How were STAR*D participants identified and from where were they recruited? Where did they receive treatment?
- How many participants were in the study? How did participants move through the stages of the trial? Describe **Figure 1**.
- How did randomization in STAR*D work? What role did participants have in choosing treatments? What do you think of this aspect of the study design?
- What medications were studied? Do you agree with the choice of medications?
- How were response and remission defined? Why do you think these categories were chosen as outcomes, rather than having a continuous outcome such as HRSD score?
- What is meant by "measurement-based care (MBC)"? How was MBC used in STAR*D?

Results:

- What did STAR*D tell us about the effectiveness of treatment for depression?
- What happened in level 1? How did participants treated in psychiatry clinics compare to those treated in primary care? Were there aspects of the study design that may have influenced this?
- On pg 452, the authors state: "As most participants elected to allow randomization to switch or augment strategies (not both), the study was not adequately powered to compare outcomes for switch versus augment treatments." What do they mean by this? Given this piece of information, did you think it was the correct study design choice to allow participants some ability to choose which treatments they would be randomized to?

- Summarize the results of level 2 for both switch and augmentation strategies.
- What were the findings regarding CBT? Do you think STAR*D was able to compare medications to CBT? Were you surprised by the number of participants willing to be randomized to CBT?
- How big was the sample size for level 3? Which participants entered this portion of the study?
- What were the level 3 treatment strategies? Do these seem reasonable to you? Do you think any other treatment options should have been included?
- What proportion of participants achieved response and remission in level 3? How does that compare with response/remission rates in levels 1 and 2? How do you understand this? What about level 4?
- In summarizing the results across the STAR*D levels, the authors report that, while the treatment response rate went down across the levels, the treatment intolerance rate went up. How do you understand this? Do you think the study design influenced this?
- What happened to the participants who entered long-term follow-up? What (if anything) does this tell us about the natural history of depression?

Conclusion

- At the end of the article, the authors make a series of conclusions. Do you agree or disagree with the following statements? Why or why not?
    - "The overall attrition from the study at all levels of treatment indicates a need to institute preventative procedures involving patient education and attrition-monitoring approaches for all patients."
    - "Our findings of minimal differences in clinical presentation between primary care and specialty care patients supports the use of the same methods for screening and measuring treatment outcomes in both settings."
    - "Using objective measurements of symptoms and side effects may be helpful when making adequate dosing and time frame determinations to maximize symptom reduction and minimize side effects."
    - "In the context of acceptable side effects, clinicians may want to consider at least 8 weeks of treatment before making a treatment change due to lack of efficacy."
- What are the implications of STAR*D for clinical practice?

Further reading on STAR*D

Primary outcomes papers:

1. Trivedi MH et al. (2006) Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *AJP* 163:28-40.
2. Rush AJ et al. (2006) Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *NEJM* 354:1231-1242

3. Trivedi MH et al. (2006) Medication augmentation after the failure of SSRIs for depression. *NEJM* 354:1243-1252.
4. Thase et al. (2007) Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR*D report. *AJP* 164:739-752.

Post-Guide (PGY2-4)

Warden D. *et al*. 2007. The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.

Article Summary

This article is a nice summary of the STAR*D trial, the largest effectiveness study ever done on treatment of Major Depression. STAR*D was divided into four stages, or levels. Subjects were all treated the same way in level 1, then progressed through the next level if they were still depressed at the end of each level. The entry criteria were broad and designed to select typical patients from the "real world," both primary care and psychiatric settings from all over the country. The levels were structured from "safest" to "highest risk" and so that subjects had the opportunity to express preferences for some treatments over others, in order to study treatment acceptability.

Since the paper does a great job of describing the levels (see figure 1) I won't do that again here. Instead I'll try to focus on the most important results. Most broadly, the important result of level 1 is that the "worse" a person was at the beginning of the study, the less likely they were to remit. This was true regardless of the type of clinical "hit" – whether a comorbidity such as anxiety, a social stressor or a disease characteristic such as chronicity. This seems like common sense, but STAR*D established the need to ask more questions about treatment of depression because it showed that we're best at treating those who weren't that badly off to begin with.

In level 2, the most important conclusion was that you should not continue patients on medication that isn't working, or is only partially working, but that it doesn't really matter what kind of change you make, both switching and augmentation arms had equivalent results. The results of subject choice of randomization options are also pretty interesting – did it surprise you that only 1% of subjects were willing to be randomized to all the options? One thing that surprised the research team was the low acceptability of cognitive therapy – only 29% of subjects were willing to consider it.

The first thing to note about level 3 is that the number of subjects has dropped considerably now, to less than 400, still a good trial size but becoming small as subjects are split into groups. Level three begins to include drugs that are definitely not first line treatments. Many definitions of treatment resistant depression now take the "third trial" a cut off point, based on the fact that rates of remission dropped off at level three, after being about the same in levels 1 and 2. Looking at figure 2, the level three arm that "looks" the best is clearly T3 augmentation, but there were still no significant differences between arms in level 3.

Finally, only about 100 subjects were left in level 4, which compared tranylcypromide (an MAOI – I've never even used it, have you?) and a venlafaxine/mirtazapine combination. The tolerability of the MAOI was low and caused further drop-out during the trial. Both arms had minimal remission rates, suggesting that when treating patients who have many medication failures, taking side effects into account (i.e. not making things even worse for the patient) may become more important.

Figure 4 shows important results on relapse rates. Subjects were followed for 12 months from the end of the last level they completed, but the mean time to relapse was much less than a year.

Comments

STAR*D was concerned with how patients reach remission. Drs. Rush, Trivedi and others in the depression center had previously published extensively on achieving functional recovery from depression and concluded that the rates of relapse were proportional to the level of symptoms remaining after treatment of any episode. Therefore reducing symptoms as much as possible was seen as critical in the design of STAR*D. Similarly there was interest in factors that are often ignored in clinical research such as the willingness of patients to accept treatment(s) and biopsychosocial factors that might effect outcome. The results clearly support the idea that patients who achieve remission have lower relapse rates, and that relapse rates increase the more trials it took to achieve at least response. They also clearly show that the more burden of illness a patient has, the worse their outcome will be.

Overall the results support more assertive treatment of depression than patient often receive, especially in primary care settings, with remission as a goal. However, the high rates of drop-out – highest among the poorest and sickest patients – suggest that even the modest results of STAR*D may be optimistic when it comes to treatment of depression.

Pre-Guide (Intern)

Warden D. *et al*. 2007. The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.

Sox H. and Lewis R. 2016. Pragmatic Trials: Practical Answers to “Real World” Questions. *JAMA* 316 (11) 1205-1206.

Reasons for choosing this research article

- STAR*D, the Sequenced Treatment Alternatives to Relieve Depression trial, is a major effectiveness trial of antidepressants. The trial had many steps and parts, and this review article provides a summary of STAR*D outcomes.
- By reading this review article, we can look at STAR*D comprehensively and consider how well the trial met its goals.
- The STAR*D project was lead by Dr. John Rush and Dr. Madhukar Trivedi, both of whom were psychiatry faculty at UTSW during the project.

Background:

- What were the goals of STAR*D? What clinical scenarios did STAR*D attempt to clarify?
- The S of STAR*D stands for “sequenced.” What does this describe?

Methods:

- How were STAR*D participants identified and from where were they recruited? Where did they receive treatment?
- How many participants were in the study? How did participants move through the stages of the trial? Describe **Figure 1**.
- How did randomization in STAR*D work? What role did participants have in choosing treatments? What do you think of this aspect of the study design?
- What medications were studied? What do you think about the choice of medications?
- How were response and remission defined? Why do you think these categories were chosen as outcomes, rather than having a continuous outcome such as HRSD score?
- What is meant by “measurement-based care (MBC)”? How was MBC used in STAR*D?

Results:

- What did STAR*D tell us about the effectiveness of treatment for depression?
- What happened in level 1? How did participants treated in psychiatry clinics compare to those treated in primary care? Were there aspects of the study design that may have influenced this?

- On pg 452, the authors state: “As most participants elected to allow randomization to switch or augment strategies (not both), the study was not adequately powered to compare outcomes for switch versus augment treatments.” What do they mean by this? Given this piece of information, did you think it was the correct study design choice to allow participants some ability to choose which treatments they would be randomized to?
- Summarize the results of level 2 for both switch and augmentation strategies.
- What were the findings regarding CBT? Do you think STAR*D was able to compare medications to CBT?
- How big was the sample size for level 3? Which participants entered this portion of the study?
- What were the level 3 treatment strategies? Do these seem reasonable to you?
- What proportion of participants achieved response and remission in level 3? How does that compare with response/remission rates in levels 1 and 2? How do you understand this? What about level 4?
- In summarizing the results across the STAR*D levels, the authors report that, while the treatment response rate went down across the levels, the treatment intolerance rate went up. How do you understand this? Do you think the study design influenced this?
- What happened to the participants who entered long-term follow-up? What (if anything) does this tell us about the natural history of depression?

Conclusion

- Consider now the JAMA article:
    - Contrast internal and external validity
    - Do the criteria for pragmatic trials proposed by Tunis et al apply to STAR*D?
    - Why was a pragmatic trial design used in STAR*D?
    - Do you agree or disagree with the following statements? Why or why not?
    - “Using objective measurements of symptoms and side effects may be helpful when making adequate dosing and time frame determinations to maximize symptom reduction and minimize side effects.”
    - “In the context of acceptable side effects, clinicians may want to consider at least 8 weeks of treatment before making a treatment change due to lack of efficacy.”
- What are the implications of STAR*D for clinical practice?

Further reading on STAR*D

Primary outcomes papers:

5. Trivedi MH et al. (2006) Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *AJP* 163:28-40.

6. Rush AJ et al. (2006) Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *NEJM* 354:1231-1242
7. Trivedi MH et al. (2006) Medication augmentation after the failure of SSRIs for depression. *NEJM* 354:1243-1252.
8. Thase et al. (2007) Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR*D report. *AJP* 164:739-752.

Post-Guide (Intern)

D Warden *et al*. (2007) The STAR*D Project Results: A Comprehensive Review of Findings. *Current Psychiatry Reports* 9:449-459.

Take Home Summary

The Sequenced Treatment Alternatives to Relieve Depression trial (STAR*D) was an attempt to comprehensively study the treatment of major depression in real world, outpatient settings. The inclusion criteria were broad to include the majority of patients with non-psychotic major depression who were appropriate for outpatient treatment; patients with psychiatric and medical comorbidity, including substance use disorders, were included. Participants were recruited from both primary care and psychiatric specialty care, as depression is a common illness in both of these settings.

In the first stage of STAR*D, all study participants (n=2876) received treatment with citalopram, which was chosen as a prototypical SSRI. About 1/3 of these participants achieved remission of depression and entered a naturalistic follow up phase, in which their treating physician continued to treat their depression according to clinical practice. The 2/3 of patients who did not achieve remission had the option to enter level 2.

In level 2, patients were randomly assigned to one of the following treatments: switch medication from citalopram to bupropion-SR (the approved formulation at the time of the study), sertraline, or venlafaxine-XR; switch from citalopram to cognitive therapy or have cognitive therapy added to citalopram treatment; or augment citalopram treatment with either bupropion-SR or buspirone. A critical feature of the STAR*D design was that participants could choose which interventions they were willing to be randomized to (e.g., a patient could say they were willing to be randomized to any medication augmentation strategy but refuse any treatments that required switching medications). The authors chose this design to mimic clinical practice, in which patients have input on the treatments they receive. This decision had some unexpected consequences, as an uneven number of patients chose augmentation vs switch, making it statistically impossible to directly compare these strategies. Additionally, relatively few participants accepted randomization to cognitive therapy, limiting the ability to draw conclusions regarding the efficacy of this strategy, particularly in comparison to medication. Surprisingly, despite all of the differences between the different treatment strategies, the rate of remission for each intervention was similar. About 25% of patients who were randomized to a medication switch achieved remission, regardless of whether they switched to venlafaxine, bupropion, or sertraline. Among those who received augmentation, remission and tolerability was slightly better for augmentation with bupropion than augmentation with buspirone. Thus, for patients who have not responded to treatment with a single SSRI and want to switch medication, switching to a different SSRI, an SNRI, or bupropion are equally good choices; augmenting the SSRI with bupropion is somewhat better adding buspirone for patients who want to continue on the original SSRI.

As participants progressed through the levels of STAR*D, treatment outcomes got worse. Participants who did not achieve remission in level 2 moved on to level 3, in which the treatment options were switch to mirtazapine, switch to nortriptyline, augment current agent with lithium, or augment current agent with T3. Fewer than 20% of patients who were switched to mirtazapine or nortriptyline achieved remission, and around 20% of patients who were augmented with lithium or T3 achieved remission, with no differences between treatments within the switch group or within the augmentation group.

Those who did not achieve remission in level 3 had the option to continue to level 4, which compared switch to trancylpromine (an MAOI) or switch to the combination of mirtazapine and venlafaxine. At this point, only about 13% of patients achieved remission, with no differences between groups.

STAR*D provided further evidence that measurement based care—systemically measuring a patient's symptoms and adjusting medications based on symptom level—is effective in the treatment of depression. Treatment rates among patients receiving care in primary care and psychiatry were the same, demonstrating that primary care providers can be effective at treating depression. STAR*D also demonstrated that patients do better over the long-term when they achieve remission of symptoms.

This article is unique among those we read in journal club in that this article is a review of several other articles rather a primary report of findings. This article was chosen because there is no single STAR*D outcomes paper, as the results of different levels of STAR*D were published separately.

Pre-Guide

RD Weiss et al (2011). Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. Arch Gen Psych 68(12): 1238-1246.

Reasons for choosing this article

- Treatment for prescription opioid dependence is an important clinical issue that many psychiatrists will see in practice. This article has important information about the medical management of patients with prescription opioid dependence.
- The authors used an unusual study design, and it's worth discussing the pros and cons of this particular design.

Background

- What is your clinical experience with seeing patients with prescription opioid use disorders? In what settings have you worked with these patients?
- The authors reports findings from previous studies about the ways in which patients with prescription opioid use disorders differ from those with heroin use disorders. What are these differences? Would you expect these differences to affect treatment?
- What (if any) hypothesis do the authors have for the study?

Methods

- The authors describe the study as a "randomized, 2-phase, adaptive treatment research design" and "a sequential multiple-assignment randomized trial." In your own words, what do these phrases mean? How does this design fit with the goals of the study?
- 635 patients were randomized--do you consider this a small, medium, or large trial?
- What do you think of the inclusion and exclusion criteria? How well do the participants in the trial match patients you would expect to see in clinical practice?
- How did the authors define "successful outcome" for phase 1 vs phase 2? Does this seem reasonable? How would you define a "successful outcome" for a patient with prescription opioid dependence?
- For analysis, the authors used a common convention in addiction trials: "missing urine samples were considered positive for opioid use." In your own words, what does this mean? How might it affect the outcomes?

A technical point from the methods:

- Why were study participants required to be in withdrawal (>8 on Clinical Opiate Withdrawal Scale) for induction on buprenorphine-naloxone?

Results

- Looking at Table 1--what do you think of the characteristics of the sample? How does this match with the patients you've treated/expect to treat?
- How did Phase 1 of the study go?
- What about phase 2?
- In both phases, how did patients do while on buprenorphine-naloxone compared to after the medication was tapered?
- In each phase, what did counseling add to the medication effects?
- In Table 3, the OR=10.6 for phase 2 end of treatment. In your own words, what does this mean?
- Were there any patient characteristics (e.g., pain, history of heroin use) that affected how well participants did?

Discussion

- What do you take away from this study?
- What does this study tell us about the use of buprenorphine-naloxone and counseling for prescription opioid use disorders?
- What are your thoughts regarding buprenorphine-naloxone treatment vs taper of treatment? In what circumstances might you recommend one vs the other? How would describe the evidence for your choice to a patient?
- This study was conducted about 10 years ago, and it was published 6 years ago. Do you think this study remains relevant to current practice?

Post-Guide

RD Weiss et al (2011). Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence. Arch Gen Psych 68(12): 1238-1246.

Article Summary

This article addresses two very important issues in the treatment of opioid use disorders. First, whether prescription opioid--rather than opiate (heroin) addiction--can be treated with shorter duration of therapy, and second, whether adding addiction specific counselling improves outcomes. Sadly, it should not be necessary to explain the importance of improving treatment for opioid use disorders in the contemporary United States. The hope that opioid users would be easier to treat than users of heroin may have led to under-treatment, making studies like this critical in mobilizing services for opioid use and in motivating doctors to prescribe less of these drugs.

This study chose a fairly select group of opioid users, excluding those with significant heroin use and those with significant other psychiatric comorbidity and chronic pain requiring ongoing management with opioids. As the authors point out, the included participants were generally considered “favorable” for recovery because of their relatively uncomplicated clinical presentation and better socioeconomic status. Therefore, they represented an excellent sample to test short term treatment. The therapy of choice in this study is buprenorphine-naloxone with or without opioid dependence counselling. The subjects received relatively intensive medical management that included weekly visits.

Similar to CATIE or STAR*D, all the subjects received treatment and went through two stages – in the first they received two weeks of buprenorphine/naloxone before being tapered. If during that phase at any time their use exceeded the threshold for response set by the investigators, they immediately went to stage two in which the treatment lasted 12 weeks instead of two. Use was measured via urine drug screens as well as self-reports. Participants had what might seem like a fairly low bar to pass to be considered to respond to treatment. They could have isolated (but not consecutive) positive urine tests, and had to report no more than 4 days a month of opioid use.

The study delivered definitive though disheartening results. Only 6.6% of subjects met the criteria for success over eight weeks after the 2 week buprenorphine-naloxone period (phase 1). Counselling made no difference. With 12 weeks of medication only about 50% of subjects met criteria for successful treatment and this number plummeted back to less than 10% after 8 weeks off medication (phase 2), whether or not subjects received counselling. The results can be summed up as follows: buprenorphine-naloxone is effective treatment for many patients while they take it, but the efficacy is lost when the medication is stopped. The ideal length of treatment remains unclear, beyond this demonstration that short term treatment is not effective.

Technical Point

The pharmacodynamics of buprenorphine-naloxone are complicated. Buprenorphine is the therapeutic agent and acts primarily at mu opioid receptors (more about naloxone below). You may recall from medical biochemistry that drug interaction with a receptor has three important features 1) the site 2) the binding strength and 3) the activity. If a drug binds to the same site on the receptor as the native ligand, only one can be bound to the receptor at a time; this is called competitive binding. How well a drug competes with the native ligand for a receptor is determined by the binding strength; if the drugs binds more tightly it will displace the native ligand from the receptor. Finally, once drug is bound to the receptor it may simply sit there and prevent the receptor from being activated, or activate the receptor either more, the same, or less than the ligand. Mu opioid receptors are g-protein coupled, which you can think of as "metabolic" for the neuron (the other major type we consider in psychopharmacology being ion-channel receptors), and cause less GABA to be released. GABA is the brain's primary inhibitory neurotransmitter, so inhibiting GABA causes more neuronal firing. Because opioid receptors are expressed on only a small subset of neurons it's difficult to tie this directly to the clinical effects of opioids, so we'll just focus on the activity at the receptor.

Buprenorphine has very high affinity for the mu opioid receptor where it binds competitively with both native opioids (endorphins) and other opioid drugs including heroin. Once bound to the receptor buprenorphine causes partial activation of the receptor, much less than most other opioids. This binding pattern causes the important clinical qualities of buprenorphine treatment. It prevents the activity of other opioids by displacing them from the mu receptor while still causing some receptor activation. Patients with opioid dependence have mu opioid receptors that transmit less signal than normal, so partial activation helps put the range of downstream activity into the normal range. Thus, the psychological and physiological effects of opioid tolerance and withdrawal (including cravings) are mitigated. However, because buprenorphine displaces other opioids from the receptors but provides less activity, it can cause acute withdrawal if a patient has recently used. To prevent this, clinicians must assess patients to rule out recent use and determine the point at which withdrawal symptoms are severe enough that buprenorphine will effectively treat withdrawal symptoms. This is why it's necessary to use a threshold on a withdrawal scale to determine when it's safe to start treatment. As a rule of thumb, buprenorphine is typically started in moderate withdrawal, when the COWS score is 8-10.

You may be familiar with naloxone as an antidote for opioid overdose. It's included in this oral formulation to prevent abuse. Naloxone has almost no effect when taken orally. But if a patient tries to abuse the combo drug by injecting it, the naloxone binds even more tightly to receptors than buprenorphine, preventing it from having an effect, and, in theory, causing a patient to go into withdrawal.

Pre-Guide

Wierenga, CE *et al*. (2015) Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biological Psychiatry* (77):642-652.

Reasons for choosing this article

- This study addresses eating disorders, which many residents are interested in learning about.
- The study allows us to consider strengths and weaknesses of human laboratory and imaging studies.
- This article allows for a discussion of the neural processes for decision making, which have wide implications for many psychiatric diseases.

Background

- The authors discuss 2 brain systems, one for reward valuation and one for cognitive control. What is the proposed role of these systems in decision making? How are they thought to work together?
- Why do the authors choose to study a monetary reward instead of a food reward? Do you agree with this decision?
- How is hunger thought to affect decision making?
- In your own words, what is the hypothesis of the study?

Methods

- Who were the study participants?
- In what ways were the testing conditions standardized?
- What task did the participants perform? How did the task work? How was the task related to compensation for trial participation?

A technical point from the results:

- What is meant by the statistical term "interaction?" In Figure 3, where can you see the interaction in the graphs?

Results

- Looking at Table 1, what were the baseline differences between the groups?
- Were there differences between groups on the delay discounting task? How was performance affected by hunger vs satiety in each group? (see Figure 2)

- The authors state that “satiety differentially modulated cognitive control response by group during intertemporal choice across all trials.” What does this mean? Describe the data that support this statement.

Discussion

- What are the major conclusions of the study?
- How do these results extend what was already known?
- The authors argue that these results have implications for reward processing and decision making in substance use disorders and obesity—speculate on this.
- What do you make of the decision to compare women with *remitted* anorexia to controls? Does this strengthen or weaken the study?
- Are their treatment implications of these findings?
- What do these results tell us about recovery from anorexia nervosa?

Post-Guide

CE Wierenga *et al*. (2015). Hunger Does Not Motivate Reward in Women Remitted from Anorexia Nervosa. *Biological Psychiatry* 77:642-652.

Take Home Summary

This article describes a small human laboratory study of the neural processing of monetary reward in women remitted from anorexia nervosa. Compared to controls, women remitted from anorexia nervosa demonstrated no differences in the processing of reward was not affected by satiety compared to hunger. This observation differed from controls, who demonstrated increased activation in the ventral striatum, dorsal caudate, and anterior when completing a reward task when hungry but activation in the ventrolateral prefrontal cortex and insula when completing the same task when sated. The study used a delayed discounting task, in which participants make a theoretical choice between a lower amount of money after a short delay or a larger amount of money after a longer delay (e.g., would you rather have $10 now or $100 in a month). The choice of a small amount with less of a delay is generally interpreted as an impulsive choice, and this type of choice is often seen in patients with substance use disorders. In contrast, patients with anorexia nervosa have previously been demonstrated to consistently choose the larger amount, even to a greater extent than healthy controls. Hunger is known to increase impulsive choices. Study participants had to be at least one year symptom free from anorexia, while controls were healthy age and weight matched women. Participants completed the study task and MRI session twice: once following a 16 hour fast and once 2 hours after a standardized breakfast. The participants in the two groups were well-matched, although women remitted from anorexia nervosa had higher rates of depression and anxiety, which was expected and consistent with known comorbidity rates. The two groups performed the delay discounting task similarly—as expected, study participants chose the larger, delayed amount of money more often as the difference between the amounts got larger. Controls made decisions faster when hungry and slower when sated; this difference in reaction time was not observed in women with remitted anorexia nervosa. The imaging findings were as described above. The authors interpret these results to indicate that, even in women who do not have clinical symptoms of anorexia, hunger does not motivate rewarded decisions, which is in contrast to the ability of hunger to motivate decisions in healthy controls. These results are striking that the observation was made among women with *remitted* anorexia, suggesting that neural processing differences remain even when symptoms have resolved.

Regarding the technical points from the pre-journal club guide: A statistical interaction occurs when the dependent variable (neural activation) has a different pattern based on the two independent variables (control vs remitted anorexia and hungry vs sated). Looking at Figure 3, in the CONTROLS, activity in the right dorsal anterior cingulate is HIGHER when HUNGRY than sated. In the REMITTED ANOREXIA group, the opposite pattern is seen: activity in the right dorsal anterior cingulate is LOWER when HUNGRY. This pattern—hunger has an effect on the control group that is the opposite of the effect in the patients with remitted anorexia—is seen in several brain areas studied and highlights the differences in neural processing between the groups.

Pre-Guide

L Wunderink et al (2013). Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 70(9):913-920.

Reasons for choosing this article

- This article raises two important questions:
    - How do we know when to STOP medications?
    - What’s the best endpoint—functional or symptom recovery?
- The length of follow up in this study is rare. Consider the importance and difficulty of this study design.

Background

- In the background, the authors emphasize two things about this study: long-term follow-up and focus on functional recovery. For patients, how important are these issues?
- What knowledge gap is the study designed to fill?
- What (if any) hypothesis do the authors have for the study?

Methods

- For clarity: the current report is a follow-up to a previously published study that examined the two-year outcomes in patients with remission of a single episode of psychosis who were randomized to either maintenance therapy (MT; keep the antipsychotic dose stable) or dose reduction/discontinuation (DR). To be eligible for the original trial (and thus this study), psychosis had to be in remission, defined as no positive symptoms for 6 months. Group assignment was random but not blind. The research clinician was in charge of maintaining the group assignment (MT vs DR) during the two years of the original study, with dose adjustments made in response to symptoms. DUP=days of untreated psychosis=number days prior to receiving the initial treatment for first episode psychosis in the first study. The original study (not required reading) is available for free through the library website: Wunderink et al (2007). *J Clin Psychiatry* 68(5): 654-661.
- The participants all had a diagnosis that fit the category “nonaffective psychosis,” but many different diagnoses fit in that category. What do think of the decision to include a somewhat heterogeneous group of patients? Why might the authors make this choice? Does this affect the way you might apply this study in approaching treatment of your patients?
- To determine “functioning,” the authors used a scale that covered 7 domains of life—how well do these domains match your ideas of “functioning”?

- How are “recovery,” “symptomatic remission,” and “functional remission” defined in this study? Do these definitions seem reasonable to you?
- What do you make of the decision to present all medication information in “haloperidol equivalents”? Would anything about your interpretation of the study be different if you had more detailed medication information?
- The original trial lasted 2 years. For this trial, patients were contacted 5 years after the original trial ended and interviewed about recent symptoms. The authors describe their design as “a 7 year follow-up.” Is that terminology fair? What attempts are made to account for events of the intervening 5 years?

A technical point from the methods:

- The authors describe using a logistic regression analysis and state that “relevant variables were entered into the regression model if bivariate analysis showed a significant association (p<0.05) with recovery, symptomatic remission, or functional remission.” What does that mean? What’s “the bivariate analysis”? What (in very general terms) are the differences between “a bivariate analysis” and a “logistic regression analysis”?

Results

- Two questions about table 1:
    - What are the clinical characteristics of the study participants? Do they fit with your idea of the typical patient with a psychotic disorder?
    - Why do the authors include information on people who weren’t included in the current study?
- Which factors were associated with recovery? Symptom remission? Functional remission?
- Were you surprised by the data on the frequency and number of relapses?
- Overall, the recent haloperidol equivalent dose differed between the groups by ~1.5 mg. Is that a meaningful difference?
- The authors perform an additional analysis, in which instead of looking at outcomes based on treatment assignment in the original study, they look at outcomes based on the treatment the patient actually received (“as-treated post hoc comparison,” pg 918).What do you make of the significant difference in the number of relapses in this comparison compared to the similarity in relapse rate in the main analysis?

Discussion

- What do you take away from this study?
- In the abstract of the original study, the authors state “only a limited number of patients can be successfully discontinued.” The authors describe their current results as “identify[ing] major

advantages of a DR [dose reduction/discontinuation] strategy." What are the advantages of the DR strategy? How do you make sense of these opposing statements from the different time points? In other words, what are the risks and benefits of the dose reduction strategy?

- The authors state "the major issue is, of course, whether these striking results may be attributed to the treatment strategies in the original trial." Well, can they?
- Speculate on the "psychological impact of having been in the DR strategy" (pg 919).
- What might be the mechanism by which better functional recovery is observed in patients treated with lower doses of antipsychotics?

Post-Guide

L Wunderink et al (2013). Recovery In Remitted First Episode Psychosis At 7 Years Of Follow-Up Of An Early Dose Reduction/Discontinuation Or Maintenance Treatment Strategy: Long-Term Follow-Up Of A 2-Year Randomized Clinical Trial. *JAMA Psychiatry* 70(9):913-920.

Take Home Summary

This study addresses the important issue of how to best treat patients with psychotic disorders. Starting in the early 1990s, Nancy Andreasen and colleagues began to study patients with first episode psychosis and their response to treatment. Using novel-at-the-time imaging methods they suggested a profound and disturbing idea: antipsychotics may cause atrophy of some brain regions and potentially worsen some aspects of psychotic disorders or patient functioning. Culturally this coincided with the return of many chronically psychotic patients to care in the community due to the closure of state hospital facilities and the corresponding increase in advocacy for persons living with severe mental illness. Although much prior research suggested that patients required consistent "adequately dosed" antipsychotic medication, patient advocates who promoted the idea of *recovery* believed that antipsychotics were not always helpful in achieving this goal. These two threads led to clinical investigation of a wider range of effect of antipsychotics. The present article seeks to further explore these ideas and carefully study the outcomes important to functional recovery in patients treated with antipsychotics for first episode psychosis. One important aspect of this for this publication is the length of follow-up. This study followed patients for an unprecedented seven year period to examine outcomes.

Recovery implies both *symptomatic remission* (presence or absence of psychosis) and *functional remission* (self-care, family relationships, vocational functioning). Therefore a combination of these two was measured as the primary outcome. Young adults experiencing their first episodes of psychosis were randomly divided into two treatment groups – some received maintenance therapy (MT) after responding to an initial course of antipsychotics, while others were assigned to dose reduction (DR) in which a taper of drug was attempted. In the original study, outcomes were followed for two years; here data from a follow up assessment five years after the original study ended was performed. During those 5 years, the study participants were treated by psychiatrists not associated with the trial. Overall recovery rates (symptomatic and functional recovery for at least 6 mo at the time of assessment) were doubled (40% vs 18%) in the patient originally randomized to DR vs MT at the seven year follow up point. The rate of *symptomatic* remission did not differ between groups, while the rate of *functional* remission was 46% in the original DR group compared to 20% in the original MT group. Although the rate of relapse of psychosis was higher among the DR group in the first two years, the relapse rate was similar between groups after 7 years. A logistic regression analysis identified being assigned to the DR group in the original study as one factor significantly associated with recovery and functional remission.

These results suggest that less antipsychotic use is associated with more recovery, in line with the theory that antipsychotics may have harmful effects outside of the traditional side effects. The authors mention this possible impairment as well as the psychological impact of original treatment assignment (patients felt more in control of their treatment or felt that they were not as impaired by their illness because their meds were tapered), or differences in illness course during the intervening 5 years that were not easily assessed. Because there was a gap in data collection of five years, its possible that factors that were not identified influenced the results. The other main take-home point is that while the MT group relapsed later overall than the DR group, over time both groups looked similar. This highlights the importance of long term follow up in understanding the course of mental illness but also suggests that antipsychotics may not provide long term improvements in symptoms.

It's important to recognize that titrating medications with the goal of minimizing symptoms may not always help patients function at their highest capacity. Antipsychotics are known to treat only psychosis, not the full spectrum of functional impairment associated with psychotic disorders. The FDA approval process is biased towards symptom reduction and might have allowed drug makers to ignore functional outcomes in testing their medications for approval, whether or not actual biological harms to the brain are responsible. As a clinician, we should always consider the function of our patients in making decisions. To learn more about recovery as a concept in mental health, advocacy groups such as NAMI are very useful.

Technical Point

Regression analysis is one of the primary methods used to understand the relationships between variables in a study. In this study the authors use the term "bivariate analysis" (results of which are presented in Table 3) to refer to regression using one dependent (outcome) variable and one independent variable. Often, though, we want to know whether multiple other variables - that have relationships to each other - affected outcome. This may be because we believe these other variables may be confounding the results (for example, older people will have a longer duration of illness than younger people so age may explain an apparent relationship between outcome and illness duration). You can look at the variables in pairs and do many bivariate regressions but it is preferable to do a regression with all the potentially relevant variables at once. Logistic regression is one of the more common methods used and gives a significance value for each variable with outcome. There is often controversy about which factors to include in the analysis; here the authors picked the things that were noted to be significant when examined using bivariate analysis as "potentially relevant." The factors that remain significant in the multivariate logistic regression are thought to be those that have more effect on outcome, while those that don't are thought to be confounders. These methods are not fool proof – in this study in particular there are lots of variable that weren't even measured – but help us avoid issues of confounding and calculating too many p-values independently, potentially missing important interactions among variables.

Pre-Guide

R. Yehuda *et al*. (2014) Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 171(8):872-880.

Reasons for choosing this article

- The article raises an interesting clinical question: what does it mean for a child to have a parent survive a major trauma, even before the child is conceived? How does it differ if the mother, father, or both parents were affected?
- The article allows us to think about the nature of transmission of parental experiences to children, and the relative role of biological modification of gene expression and experiences.
- This article lets us discuss molecular biology concepts that come up in many areas of psychiatry, specifically epigenetics and DNA methylation.

Background

- What was known about the impact on the mental health of offspring whose parent(s) had PTSD? Did this differ depending on whether the mother, father, or both have PTSD?
- How is HPA-axis function believed to be affected in the offspring of parents with PTSD?
- What is the hypothesis of the study?

A technical question:

- The authors don’t discuss this, but good to review: what is DNA methylation? What is the impact of methylation on transcription? What is meant by the term “epigenetics?”

Methods

- Who were the study participants? Which participants were the experimental group and which were controls?
- What tools were used to assess the mental health of study participants?
- How were the groups combined/compared in the statistical analysis? How does this compare with the description of participants that were recruited earlier in the methods?

Results

- Were there parents who survived the Holocaust who did not have PTSD? Does this tell you anything about PTSD?

- Looking at Table 1—how would you describe the study participants? What do you notice about the numbers of participants with history of MDD or anxiety disorders? How did parental PTSD affect the occurrence of these disorders in offspring?
- What was the effect of parental PTSD on methylation of the glucocorticoid receptor promoter in offspring? How did it differ depending on the sex of the parent?
- What was the observed relationship between promoter methylation and gene expression? What were the functional consequences of methylation (as determined by the dexamethasone suppression test)?
- For Figure 2, it helps to look at it in color. Looking at the figure overall, what does the grouping together of colors tell you? What do blue and red mean (hint: read the figure legend)? How do you interpret the cluster analysis? What does it tell you about common traits in the groups sorted by parental PTSD?

Discussion

- What do you take away from this study?
- The authors give their rationale for why they think methylation of the glucocorticoid receptor promoter is important in the paragraph starting “the hypothesis that GR-1_F methylation would be associated with early adversity . . .” on page 878. What do you think of this model?
- Speculate on the mechanism—how does parental trauma lead to an effect on offspring gene regulation and expression?
- Another way of asking the question above--think about the nature vs nurture argument: do you think parents with PTSD transmit altered genes to their children, or do you think the children’s genes get altered because of the parenting style of parents with PTSD?

Post-Guide

R. Yehuda *et al*. (2014) Influences of maternal and paternal PTSD on epigenetic regulation of the glucocorticoid receptor gene in Holocaust survivor offspring. *Am J Psychiatry* 171(8):872-880.

Take Home Summary

This article observed differences in the methylation pattern of a glucocorticoid receptor and psychiatric traits among adults depending on the pattern of parental PTSD. To create a population with parental PTSD, the authors studied the children of Holocaust survivors. Participants were recruited based on parental Holocaust survivorship and then categorized into 4 groups based on whether their parents had PTSD: no parental PTSD, maternal PTSD only, paternal PTSD only, or both maternal and paternal PTSD. A small number of controls whose parents had no exposure to the Holocaust and did not have PTSD were also included. Parental PTSD was determined based on interview/questionnaires completed by study participants (i.e., adult children); study participants completed several other assessments of mood, anxiety, attachment style, and other psychiatric symptoms. Paternal PTSD only was associated with high gene methylation and less gene expression, greater childhood trauma exposure, a less secure attachment style, and greater sensitivity to violence. Having both parents with PTSD (maternal and paternal PTSD) was associated with lower methylation and higher gene expression and a sense of being affected by vicarious trauma experience (i.e., the Holocaust experience of their parents). Additionally, there was a negative association between gene methylation and cortisol response on a dexamethasone suppression test, such people with high gene methylation showed less cortisol suppression in response to dexamethasone (i.e., cortisol remained high when it should have been suppressed).

There was no effect of simply having a parent who is a Holocaust survivor if that parent did not also have PTSD. The authors suggest that the results can be understood in relation to animal studies that have demonstrated that the raising of rat pups in a high stress environment alters maternal care of the pups; the pups show increased methylation of the glucorticoid receptor gene, less glucocorticoid receptor protein expression, and an exaggerated physiological response to stress. They propose that these results suggest a similar mechanism in humans in which parental traits alter gene expression in offspring.

Regarding the technical questions from the pre-journal club guide: The term "epigenetics" refers to alterations in gene expression that are not due to changes in the DNA sequence. Most epigenetic modifications affect the chemical or physical structure of DNA without affecting the DNA sequence itself. Epigenetic modifications are involved in normal processes like X-chromosome inactivation, tissue differentiation (explains how every cell has the same DNA but different genes are expressed in different tissues), and gene imprinting as well as development of diseases. Epigenetic modifications account for the observation that the same DNA sequence can be expressed differently in different people (for example, the deletion of the same portion of chromosome 15 underlies both

Angelmann’s and Prader Willi syndrome, but the two conditions have different phenotypes). Epigenetic changes, while mostly occurring in utero and stable throughout a lifetime, can be induced by events occurring over the lifetime. DNA methylation is one type of epigenetic modification and describes the addition of a methyl group to cysteine nucleotide. This modification blocks transcription; thus, methylation of the promoter region prevents gene expression. Methylation occurs at areas termed CpG sites, in which a cysteine nucleotide is followed by a guanidine nucleotide. Other mechanisms of epigenetic modifications are (1) histone modifications that regulate which pieces of chromatin available for transcription to mRNA and (2) non-coding RNAs that interfere with mRNA translation to protein.

Pre-Guide

Yovell, Y *et al.* (2016). Ultra-low-dose buprenorphine as a time-limited treatment for severe suicidal ideation: a randomized controlled trial. *American Journal of Psychiatry* 173(5): 491-498.

Reasons for choosing this article

- This article presents a novel, medication-based approach to the treatment of acute suicidality.
- This article lets us consider the ethics of randomized controlled trials in suicidal patients.

Background

- In the population, is suicide rare or common?
- What is the connection between depression and opioids, as presented in the background?
- What gaps in the literature do the authors feel this study addresses?
- What factors were the authors attempting to balance with the dose of buprenorphine and duration of treatment in this study?

Methods

- Who were the study participants? Do you think the inclusion and exclusion criteria were appropriate? How long did it take the authors to recruit the 60 participants randomized in the study?
- How many patients were randomized to buprenorphine vs placebo? Why did the authors choose this ratio?
- What percentage of prescribed medications was actually taken by the study participants? How did the authors determine this? Was medication adherence in this study high or low?
- The authors required the following for a person to enter the study: "being in treatment with a mental health professional, clinic, or hospital that was not part of the study team. . . . . we obtained the approval and collaboration of their treating clinicians." Why was this required?
- What do the authors mean by the statement: "given the ethical considerations, the study was designed as an adjunctive trial"?

A technical point from the methods/results:

- In the section of the methods describing the statistical analysis, the authors use the term "last observation carried forward." What does this term refer to? What statistical problem is this meant to address? What are other ways of dealing with this issue?

Results

- How many of the randomized participants actually received medication (either placebo or buprenorphine)? Is this high or low? What do you make of this?
- What are the main findings of the study, as presented in Figure 1?
- What are the additional analyses presented in Figures 2 and 3? Why did the authors (or maybe reviewers?) think these analyses were important to include?
- How do you interpret Figure 4? Did buprenorphine have a similar effect on symptoms of depression as it did on suicidality?

Discussion

- What do you take away from this study?
- Do you think that low dose buprenorphine should become part of treatment for patients with suicidal ideation? Why or why not? Currently, what are limitations on putting this strategy into practice?
- What set of symptoms do the authors believe is related to the effects of buprenorphine? How might you test that in a future study?
- In discussing the limitations of the study, the authors note the heterogeneity of the study population. What do they mean by this? How does it affect interpretation of the results?
- Read the disclosure statements. Do these raise any concerns?

Post-Guide

Yovell, Y *et al.* (2016). Ultra-low-dose buprenorphine as a time-limited treatment for severe suicidal ideation: a randomized controlled trial. *American Journal of Psychiatry* 173(5): 491-498.

Summary

This article uses a fairly straightforward study design to approach a tough question in a novel way. Interest in the opioid system has dramatically increased in concert with the increase in opioid use disorders in the United States. It isn't a surprise that the same features that may make opioids addictive may have benefit for mood disorders; the conundrum is whether it is possible to separate the risks from the benefits enough to produce a drug worth prescribing to patients. One strategy used for opioids and other drugs of abuse is microdosing – giving a drug in amounts typically an order of magnitude below those usually employed. Here the authors apply this strategy to patients with suicidal ideation (largely in the context of difficult to treat borderline personality disorder (BPD)) using buprenorphine, a mu and kappa opioid receptor acting drug typically used to prevent relapse in patients with opioid addiction. The authors give a great historical and scientific summary behind the rational for this choice, so I won't repeat that here. Once the foundation for the project is established the authors spend a lot of time addressing the logistics of managing risk and the unreliable nature of such seriously ill patients.

The authors carefully screen patients to rule out substance use disorders and then randomize them in a 2:1 ratio to buprenorphine:placebo. Although from a statistical perspective unequal group sizes are undesirable, in studies of very severely ill patients there is usually a preference to avoid the effort and risks of study participation in a placebo group, so it is common to skew the group ratios as is done here. Because of the theoretical suggestion that abandonment distress in BPD may be specifically targeted by buprenorphine, subjects were also screened for this disorder. The subjects return for weekly visits, both to keep close monitoring of symptoms and to minimize the amount of medication subjects take home with them (so that they never have enough to overdose with). The primary outcome was decrease in suicidal ideation on the Beck Suicide Scale (which is not often used in the US but is a validated instrument) as well as total depression severity.

The study lasted four weeks with a one-week wash-out period to assess for symptom rebound and withdrawal symptoms. Unsurprisingly for the population, they had a high drop-out rate early in the study. While many studies of depressed patients have drop-out rates this high over an 8 or 12 week study, it is relatively unusual for almost 30% of subjects to drop out in the first week. Despite this they found a clear effect of buprenorphine on suicidal ideation. This effect was not changed by antidepressant medication but was significantly moderated by presences of BPD as a diagnosis. Patients with BPD had a very low placebo response, increasing the difference between the placebo and drug groups. They also found evidence that this effect was not purely related to a decrease in overall depression symptoms although depression did improve over the course of the study.

The results of the study support the hypothesis of the investigators and suggest that buprenorphine may have a role in treating severe or chronic suicidal ideation, especially in the setting of BPD. Given the risks of buprenorphine and the relative susceptibility of patients with BPD to substance use disorders and overdose attempts, however, much work remains to be done to improve the safety of providing such treatment to patients.

Technical point from the pre-guide:

As mentioned above, subject drop-out is a big issue in clinical research. Allowing patients to exit a study is an ethical requirement, but creates problems for analysis in two ways. First, a smaller sample decreases the statistical power of an analysis. In a study like this, which would be considered relatively high risk, often samples are thought to be best at the minimal necessary size for ethical reasons, which leaves little cushion for drop-outs. Secondly, drop-outs are not random and may undermine a randomization scheme designed to balance the two groups. Often the factors determining the drop-out pattern are not obvious or even measurable at all, but you can imagine that if many patients dropped out of the drug group for a particular side effect (say fatigue which prevented them from showing up to their appointments) then the results could be skewed, as the remaining patients in the drug group would no longer represent the full scope of the drug's action.

One method of addressing drop-outs would be to not address them! That is, to include only the subjects who finished the study. As suggested above there are major issues in using this method, so it is frowned upon. You will notice that the authors reported such results in this paper but only after reporting on results using a more active way of addressing drop-outs. This type of analysis is sometimes described as a "per protocol" (i.e., participants who completed the study as specified in the protocol) or a "completer" analysis.

Generally, such methods for addressing drop-outs (or even missed study visits, where data were not collected at a certain time point because the participant did not attend that study visit) involve "imputation" which means putting numbers into spots in the data set that are blank, using a method to predict what those numbers might have been. The simplest of these is to simply repeat the last available number in every time point left empty. So, if at baseline the subject had a score of 20 on the Beck Suicide Scale, 20 would be entered for the score at week 1, week 2, etc. This is called "last observation carried forward." This avoids a biased prediction in terms of increase or decrease in score but does not provide much else. Given that most subjects will have other factors affecting them over the study period in real life it's not reasonable to predict that scores remain stable. Still because the method is simple and makes few assumptions you will commonly see it used to preserve power and avoid bias from non-random drop-out.