FEASIBLE EVIDENCE-BASED PRACTICE IN CHILD AND ADOLESCENT MENTAL HEALTH

John Hamilton

Katsushika Hokusai

John Hamilton MD, MSc

Teaching Associate in Child and Adolescent Psychiatry, Cambridge Health Alliance, Harvard Medical School, Cambridge, Massachusetts, USA

Conflict of interest: none reported.

Suggested citation: Hamilton J. Feasible evidence-based practice in child and adolescent mental health. In Rey JM, Martin A (eds), *JM Rey's IACAPAP e-Textbook of Child and Adolescent Mental Health*. Geneva: International Association for Child and Adolescent Psychiatry and Allied Professions 2021.

This chapter shows how to improve the alignment between what you do in your practice and good evidence. Consider Alex. Fourteen-years-old, he is attending high school and tried to hang himself at home with his belt. His brother heard the labored breathing and ran into his room. He was hospitalized for a few days, but a week later was discharged and came to see you. What are feasible treatments which align with reliable, low-bias evidence?

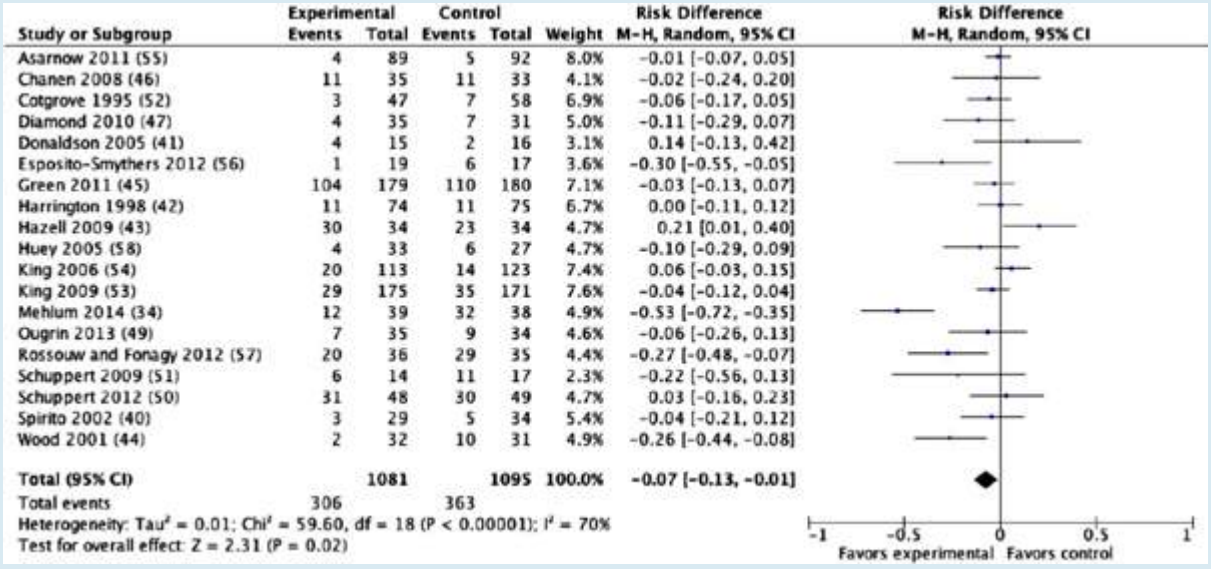
In choosing treatment, this chapter will emphasize the usefulness of *going directly to aggregated evidence from multiple randomized controlled trials*. There are several ways to make sense quickly of aggregated data, including meta-analyses and systematic reviews, with their highly useful visual summaries of results. For Alex's condition, a search online for such aggregated evidence about treatment for adolescents who attempt suicide brings up Ougrin's and colleagues' work from 2015. They offer a visual summary—often a great place to start—of their conclusions after examining 19 of the most relevant studies. Their visual summary below (Figure A.6.1) condenses results comparing intervention groups to control groups for self-harm, defined as both non-suicidal self-injury, such as cutting without intent to die, as well as suicide attempts. The overall therapeutic result is quite modest with a number needed to treat (NNT) of 14: fourteen youth need to be treated in order to prevent a single episode of self-harm (either a suicide attempt or a non-suicidal self-injury).

We have two sorts of evidence now. First, we know considerable detail about Alex from a clinical interview regards his family, his development, current symptoms and their context. Second, we have Ougrin's summary of 19 intervention studies dealing with adolescents who harm themselves. And finally, most of us

- Do you have questions?
- Comments?

Click here to go to the Textbook's Facebook page to share your views about the chapter with other readers, question the authors or editor and make comments.

Figure A.6.1 A 'forest plot' is the conventional term for this plot, showing the difference in risk for self-harm or suicide in treated adolescents versus in control adolescents summarizing results from 19 different randomized controlled trials. The 'forest' comes from the image of the horizontal bars, representing the confidence intervals for the risk difference between treated and control subjects in each study. (Wouldn't 'rowers crossing the finish line' be a better description than 'forest'?). The 'black diamond,' with its tighter confidence intervals, summarizes the results overall, its width representing confidence intervals.*



*Ougrin et al, 2015; with permission.

clinicians can say what we might do at this point—we have ideas about how we might proceed. We may have seen children like Alex and successfully treated them. There are many individual characteristics of Alex as a person that appear relevant: he is sensitive, intelligent, used to play tennis often with his brother and mother, feels alienated from his father, and looks scared. He has several important friendships and is uncertain about his burgeoning sexuality. These observations are useful evidence for Alex's treatment that might well be included in a psychodynamic formulation. *To be effective, the challenge is to make use of both sorts of evidence, the meta-analytical and the highly personal, the statistical aggregates and the subtle observations of his face and behavior.* The reward is being able to improve the alignment between your work and the ever-increasing knowledge-base as experiences and studies accumulate around the world while staying attuned to Alex himself.

Feasible evidence-based practice finds aggregated evidence by going straight to 'the top of the evidence pyramid' where meta-analyses and systematic reviews reside, and then looking for a visual summary of the results. The outcome is a startling view over the 'evidence-scape'—all or nearly all relevant randomized controlled trials (RCTs) included as well as a visual summary of the primary outcome for each RCT. Using aggregated data is feasible because it is quick. A single figure such as the forest plot in figure A.6.1 contains a remarkable amount of information.

This chapter tries to be useful to clinicians arriving with a spectrum of expertise and experience in evidence-based thinking, from the beginner to intermediate and advanced. After this introduction, the chapter has a section—basics—to teach fundamental ideas from statistics and epidemiology and may well be skipped by students and practitioners who have a firm grasp of basic statistics and of terms like standard error of the mean, odds ratio, effect size, confidence intervals, number-needed-to-treat, systemic bias versus random error, dichotomous versus continuous variables, and other basic concepts. The goal is to offer clinicians from multiple backgrounds an opportunity to become familiar with fundamental statistical ideas and terms essential to understanding evidence-based practice. The 'basics' section can also be skipped by beginners in statistics and used simply as a reference as needed to understand later sections.

After a 'mental hike' up the (imaginary) evidence-based practice pyramid, practitioners enjoy the view at the 'top of the evidence pyramid' looking out over the data summarized visually in meta-analyses and systematic reviews.

Distillation and Matching

In distillation and matching, all manuals describing treatments proven to work well for a given syndrome are 'distilled' into their most common components, also called 'elements.' Therapists then use these common elements (common, because they are shared by multiple manuals). An analogy is distilling whiskey: distilling multiple bulky manuals to produce something more potent and succinct—a short list of common elements. These common elements are then 'matched' to patients with similar symptoms.

The next section is how to understand and use methods of aggregating data such as:

- Meta-analyses
- Systematic reviews
- Network meta-analyses, and
- 'Boiling down' psychosocial interventions from multiple successful manuals into those elements which make them successful (called 'distillation and matching of core components').

For busy practitioners, these data summaries from multiple randomized controlled trials into a single document, are priceless, allowing us to see in a single visual display exactly what we want to know: what works and what doesn't.

Let's turn now to fundamental concepts from statistics and epidemiology. These concepts are essential in understanding results from randomized controlled trials and aggregated results from multiple RCTs, like meta-analyses and systematic reviews. How to help Alex will be our guide.

BASICS

This section presents fundamental concepts from epidemiology and statistics that may be new to some readers and an unnecessary review for others. Consider simply visually scanning it for now. Any parts look interesting…? It can also be skipped for now and used for reference later. Do not get bogged down into reading it straight through unless it captures your interest.

What is 'Good Evidence'?

When a 12-year-old boy and his family walk in for the first time, a good therapist is in part like a good detective—looking for evidence. Does the boy look anxiously to his mother for reassurance? Does he look annoyed and would rather not be there? How do the parents respond when asked about their family histories? As the consultation progresses, does the family have an avoidant attachment style with reluctance to engage and a distant quality? Does the boy engage with the therapist when seen alone? *Subtle observations are good evidence which is essential to practice.* Staying alert to such subtle interpersonal cues is essential.

This chapter uses the term 'good evidence' in a different way, usually referring to results of treatment trials for specific disorders, especially trials that include randomization of participants into a treatment group and a control group—as well as 'blinding' of participants, raters, and the treatment team, *with the goal of minimizing bias and random error*. The 'good evidence' referred to in this chapter supplements the good evidence clinicians gather during the clinical interview.

Bias and Random Error

Bias interferes with validity—the degree to which a study measures what it intends to measure—by creating systematic (non-random) deviations from the underlying truth (Guyatt & Rennie, 2002). Bias is different from *random error*. Random error is diminished by increasing sample size, whereas bias can remain in a sample of any size. Consider the targets (a) through (d) in Figure A.6.2 where the bull's eye in the center marks the true effect. In study (a) there was little or no systematic deviation from the true effect (little or no bias), as well as little scatter (random error). In (b), there was also little bias but much random error, and in (c) much bias but little random error. In (d) there was much bias and random error (Kabai, 2021).

Finding Evidence with Low Bias and Low Random Error

The most useful evidence to guide clinical decisions is evidence with low bias and low random error. There are many ways bias can be introduced into an RCT: missing outcome data, bias in selecting what is reported (e.g., ignoring difficult side-effects), and bias in measurement of the outcome (e.g., raters of the outcome suspect that a child is on a medication rather than placebo and, therefore, change their ratings to achieve a desired result).

Figure A.6.4 shows a graphic summary of potential sources of bias reported in RCTs making up a meta-analysis of studies using CBT to treat anxiety in children done by the Cochrane group (James et al, 2015). Cochrane reviews are notoriously tough about bias, much to the chagrin of many researchers, yet both the verbal summary and the plot summarizing the statistics of this meta-analysis

Figure A.6.2 Examples of bias and of random error in data samples (Kabai, 2021).

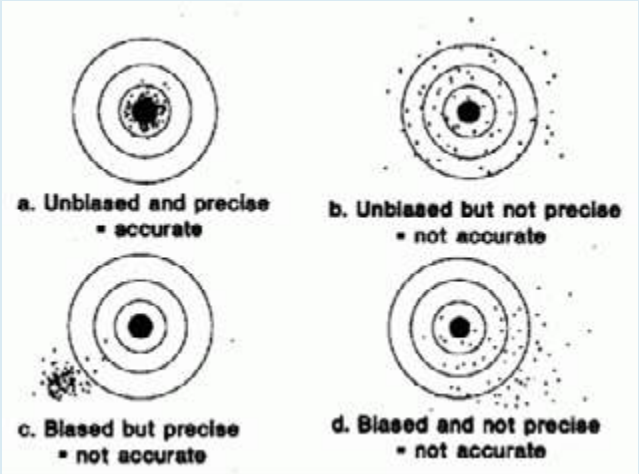


Figure A.6.3 Schematic representation of random error and bias (Kabai, 2021).

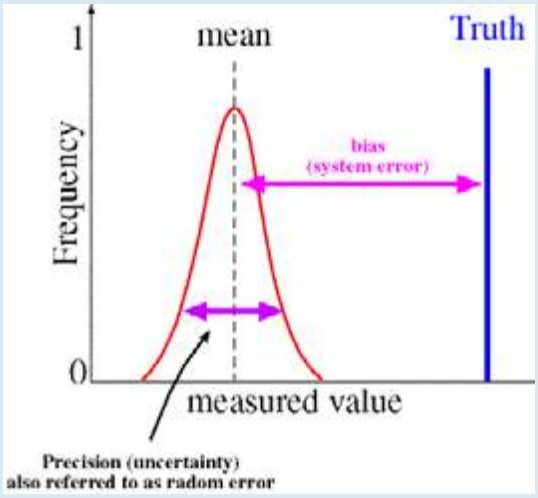
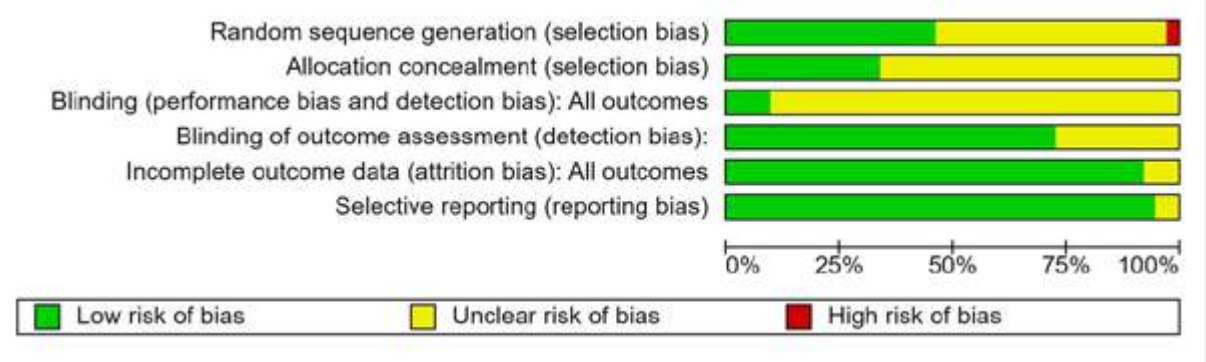


Figure A.6.4 A Cochrane review's visual summary of sources of bias in the RCTs aggregated into a systematic review of CBT interventions for childhood anxiety*



*James et al, 2015.

conclude that CBT shows a significant effect. This is not unusual. Although there are potentially significant biases in some studies (e.g., regards how well the blinding was done) the result is significant, useful, and clearly stated. While the review concludes 'the evidence is poor,' the reviewers are referring to the remaining risk of bias. Yet this Cochrane summary of evidence about the effectiveness of CBT in treating anxiety disorders in children is among the best evidence anywhere showing that CBT is an effective treatment for children with anxiety disorders. Do not be deterred from using treatments showing a significant effect in a Cochrane review even if the review concludes with statements like 'the evidence is poor.' 'Poor' is simply commenting on potential uncertainties uncovered by the extremely thorough Cochrane review processes. 'Poor' in this context means that extensive further studies done with attention to those sources of bias may achieve a different result. 'Poor' can occur in a Cochrane description of what is the best available evidence. Cochrane reviews are picky but do not be deterred.

ESSENTIAL STATISTICS FOR EVIDENCE-BASED PRACTICE

A few basic statistical concepts are all that is needed to understand many core ideas used in evidence-based approaches. First, recall the basic *normal* or Gaussian curve which describes the distribution of many aspects of nature, and which can be derived mathematically using probability theory. The most common value is the same as the ***mean***, represented below in Figure A.6.5 by the Greek letter *mu* (μ).

Standard deviation **(SD)** is a measure of variability: it reflects how much variation or scatter there is from the mean value (average). A small SD relative to the mean indicates a distribution where data points are closely clumped together (*a* and *c* in Figure A.6.2). A large SD relative to the mean indicates a distribution where data points are spread out over a large range (*b* and *d* in Figure A.6.2). SD is sometimes described as the *typical amount that cases differ from the mean.* In a 'normal' distribution of values, two out of three cases are expected to fall within plus or minus one SD of the mean.

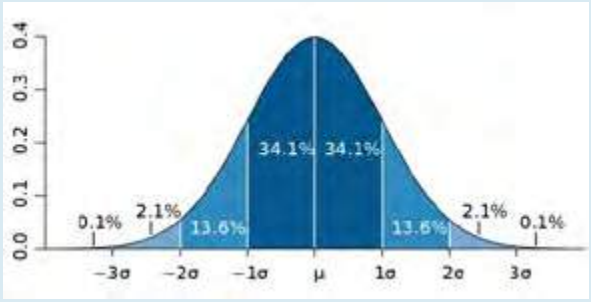


Figure A.6.5 Plot of a normal (Gaussian) distribution having a mean value represented by the Greek letter *mu* (μ) and a standard deviation, *delta* (σ), represented by each colored band. Note that slightly over one-third (34.1%) of the population of this distribution lies within one standard deviation above the mean; slightly over one-third (34.1%) also lies within one standard deviation below the mean. Only about 2% of the population lies within 2 or more standard deviations above the mean, and only about 2% lies within 2 or more standard deviations below the mean (Wikipedia).

Effect size (ES) in a research study measures how big the effect of a treatment is relative to the natural scatter in the populations studied. Effect size is also called the *standardized mean difference* (SMD) between the two groups in a study if the primary outcome variable is continuous rather than categorical. Technically, this is measured as the ratio of the difference due to treatment to the pooled standard deviation. Pooled refers to combining figures from both the control and experimental group to create a pooled standard deviation.

$$\text{Effect size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{[\text{Pooled standard deviation}]}$$

Absolute risk reduction (ARR) is a measure used to compare two different alternatives: how much does one treatment reduce the risk of a specified 'bad' outcome (like a depressed youth staying depressed) compared to either an alternative treatment or placebo? For example, in the Treatment for Adolescents with Depression Study (TADS; March et al, 2004), at 12 weeks, 65.2% of adolescents receiving placebo were not rated as improved or very much improved compared with only 39.4% of adolescents receiving fluoxetine. Therefore, the ARR of not improving for youth with major depressive disorder who took fluoxetine compared to placebo was 65.2 - 39.4 = 25.8%. In other words, fluoxetine reduced the risk of *not* improving considerably: 25.8%, or about 1 out of 4.

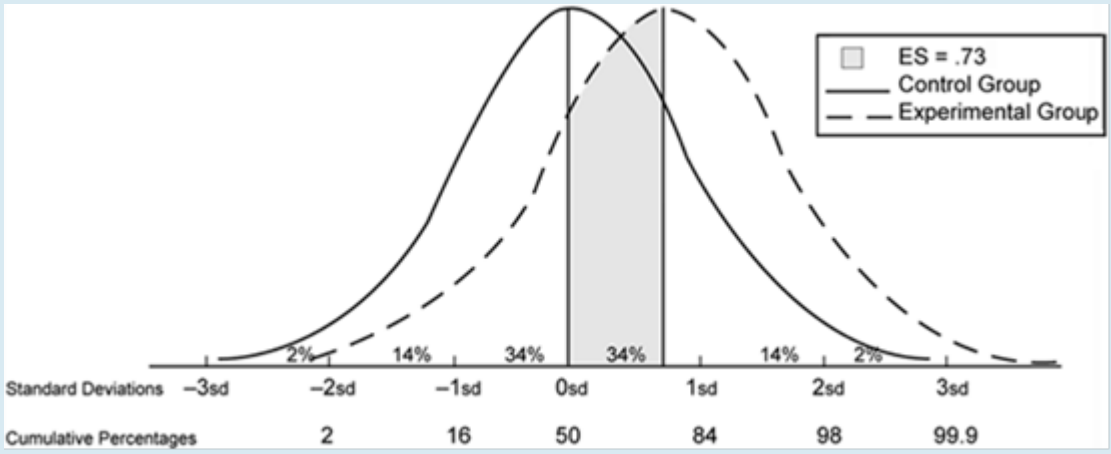


Figure A.6.6 Data from this randomized controlled trial shows that, relative to the control group, the experimental group after treatment has improved: it has moved to the right 0.73 standard deviation units. Therefore the effect size is 0.73.

Absolute risk reduction is helpful for two reasons. First, as discussed below, the ARR can be used to compute the *number needed to treat* (NNT). The formula is simply NNT= 1/ARR with the result rounded to the next whole number. Second, the ARR does not exaggerate the treatment effect as does the *relative risk reduction* (RRR) which is simply computed as the risk in the treatment group divided by the risk in the control group. In TADS, the relative risk reduction when comparing outcome for youth receiving fluoxetine to youth receiving placebo is the risk of depressive symptoms being rated as not significantly improved in the group receiving fluoxetine divided by the risk for youth not being significantly improved in the placebo group, or 39.4 / 65.2 = 0.60, generally expressed as 60%. Note that this figure is quite different from the absolute risk reduction as calculated in the paragraph above (25.8%). In general, ARR is a more useful statistic than the RRR.

Confidence interval (CI). Rather than beginning with a definition, a specific example might be more useful to gain a feel for this important concept. Suppose the minister of health tells you she is concerned about depressed teenagers and she asks you to determine the average (i.e., mean) level of symptoms of depression in 13-year-old children in the large city where you live. You accept the challenge and begin the project by arranging for 13-year-olds from all over the city to complete an inventory of depression symptoms. The mean score is 39.8 on the 49 completed questionnaires. You calculate the standard deviation and it is 9.0. The minister wants you to estimate, based on your data from the 49 returned questionnaires, how close is your mean to the *true* value—also called the population value. That is, if it were possible to assess *all* 13-year-old children in the city with this depression scale, this would presumably be very close to the *true* value. You can see, therefore, that although the *true* or *population* value is a useful concept, it is often not practical to test all the children. Yet, simply by chance you may have ended up with a higher proportion of unhappy children or of unusually happy children in your small sample than is present in the city in general, even if you managed to avoid being biased in your sampling (like choosing only students from the same school, or only students in a certain class).

You tell the minister the mean score in your sample is 39.8, as a rough estimate. She, however, wants to know more, and asks, 'How rough an estimate is that? What is the highest true value? And the lowest?' You answer, 'Yes, I can give you a range or interval but it depends on how often you want me to be right. Do you want me to be right like 80% of the time that I give you these estimates? Or 95%? Or even 99%? The more you insist that the answer I give you does contain the true answer for the population, the wider the interval will be.' The minister

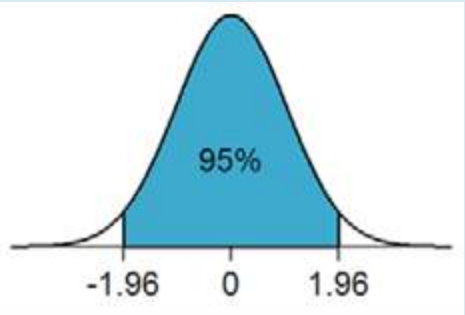


Figure A.6.7 An important fact associated with the "1.96" in the formula for calculating 95% confidence intervals: 95% of the area of a normal distribution is within 1.96 standard deviation units of the mean.

says, 'OK. I want your answers on these estimates to be correct 95% of the time.' The minister has just asked for the 95% confidence interval of your calculated mean.

A statistician can now help. She tells us that a good estimate for calculating confidence intervals in normal distributions is as follows, where N is the number of subjects generating the data (49 in this case) and SQRT means square root:

Confidence interval, upper bound = Calculated mean + 1.96 x SD/ SQRT N

Confidence interval, lower bound= Calculated mean - 1.96 x SD/ SQRT N

Therefore, the confidence intervals are:

CI_{upper} *= 39.8+1.96x9.0/SQRT 49 = 39.8+1.96x9.0/7 = 39.8+2.52 = 42.32*

CI_{lower} *= 39.8-1.96x9.0/SQRT 49 = 39.8-1.96x9.0/7 = 39.8-2.52 = 38.28*

Now you have an answer for the minister.

In summary, CIs are expressed with three numbers, the first being the mean—our best estimate of the true value. The other two numbers state the lower bound and the upper bound of the interval. The more certain we want to be, the wider the confidence interval will be. *Confidence intervals are important because statistical analysis of studies is based on the central idea that, despite making observations on a limited sample of subjects, the inferred truth to be drawn from the study will apply to the population of all such subjects.* The main purpose of a confidence interval is to indicate the precision or imprecision of the study sample as an estimate of the true population value. Confidence intervals are therefore useful and even necessary whenever an inference is made from the results of one study to the wider world. In the first example, there were only 49 children completing the depression questionnaire, but the minister hopes to generalize the findings to all 13-year-olds in the city. Also, note that even if you and the minister chose your sample wisely—e.g., from very different schools in different parts of the city to avoid bias—the problem of random variation in the level of depression symptoms remains. Merely by chance, you may have come across children with notably higher or notably lower depression levels than exist in the entire population of children of that age in the city, even though there was no significant systematic bias in your sampling method.

The ***number needed to treat***, often abbreviated NNT, can be used to summarize in a single number how effective a specific treatment is compared to placebo. NNT is defined as *the number of people we must treat in order to prevent*

How high should be the hurdle researchers require for a youth to be considered 'improved' when they calculate NNT?

Higher hurdles will raise NNTs because fewer youth will get over them.

one additional bad outcome. A bad outcome (e.g., lack of improvement) is defined by specific criteria. Therefore, low NNTs (such as 3 or 4) indicate an effective treatment, because it is common for the treatment to *convert* a patient to a successful outcome who would not have occurred without the treatment. Typical successful outcomes are improved mood or improved functioning, often defined with a cutoff point on a scale.

Note, though, that NNT depends very much on the *hurdle* that needs to be cleared to qualify as 'better' or 'cured.' *How high or low this hurdle is will impact on the NNT hugely.* When reading a paper and seeing an NNT, always note, therefore, what hurdle the authors have chosen. High NNTs (like 25, 30 or even higher) suggest either a treatment that is not very effective or an outcome that is difficult to achieve. For example, if we require that a successful outcome in a depressed adolescent is not to show a single depressive symptom, then it will be difficult to achieve this outcome and the corresponding NNT will be higher than if we had chosen a more modest outcome (e.g., a 50% reduction of depressive symptoms in a depression rating scale).

Ourgrin and colleagues (2015) conducted a meta-analysis of 19 RCTs to estimate the effectiveness of interventions for youth who self-harm. They reported that across studies 28.3 % of participants receiving the studied intervention had an episode of self-harm within 10 months (the mean of study duration) compared with 33.2 % of those receiving treatment. What is the NNT?

In this case *NNT = 1/ (0.33.2 – 0.28.3) = 1/0.0499 = 20.04*. The NNT is 21 because it is customary to round up NNT to the next whole number. 21 youth will need to be treated in order that one of them will not self-harm during the next 10 months, who would have harmed himself without treatment. However, only a few of the studies were generating almost all the effect: 3 studies had NNTs as low as 2, 4, and 5.

Number needed to harm is abbreviated NNH and summarizes how frequently a specific side-effect (or side effects) occurs with a specific treatment. NNH is defined as the number of people we must treat for a single person to be harmed by the treatment who would not have been harmed if he had received only a placebo. For example, with medications, an undesirable effect (e.g., suicidal thoughts) occurs that would not have occurred if the patient had received a placebo. Therefore, high NNHs indicate a safe treatment: many individuals must receive the treatment before a single individual is harmed by the treatment. Decisions to recommend a specific treatment will require balancing the benefits (NNT) with the risks (NNH).

Categorial vs continuous variables. The distinction between categorial and continuous variables is important because the statistical methods used to aggregate outcome from a categorical variable summarize results using odds ratios (see how to calculate odds ratio below), such as whether a child with an anxiety disorder is in remission or not. The statistical methods used to aggregate outcome from a continuous variable summarize results as standard mean differences (SMDs). This is why meta-analyses sometimes show results as odds ratios and sometimes as SMDs, depending on whether the primary outcome variable—the outcome of primary interest in the research—is categorical (right handedness) or continuous (height).

A **categorical** variable is like being left-handed or right-handed: the person is assigned to one of a limited number of categories based on some trait.

A **continuous** variable is like height or weight. The person can be anywhere along a continuum of an endless number of possible values.

Odds Ratio. To understand the term odds ratio, recall the concept of odds: the odds of an event is the ratio of the frequency (or likelihood) of its occurrence to the frequency (or likelihood) of its non-occurrence (Andrade, 2015). For example, the odds of randomly drawing an ace out of a pack of cards is 4 (since there are 4 aces in the deck) to 48 (since there are 48 other cards which you might have drawn instead). To simplify via dividing by 4, the odds of drawing an ace are therefore 1 in 12.

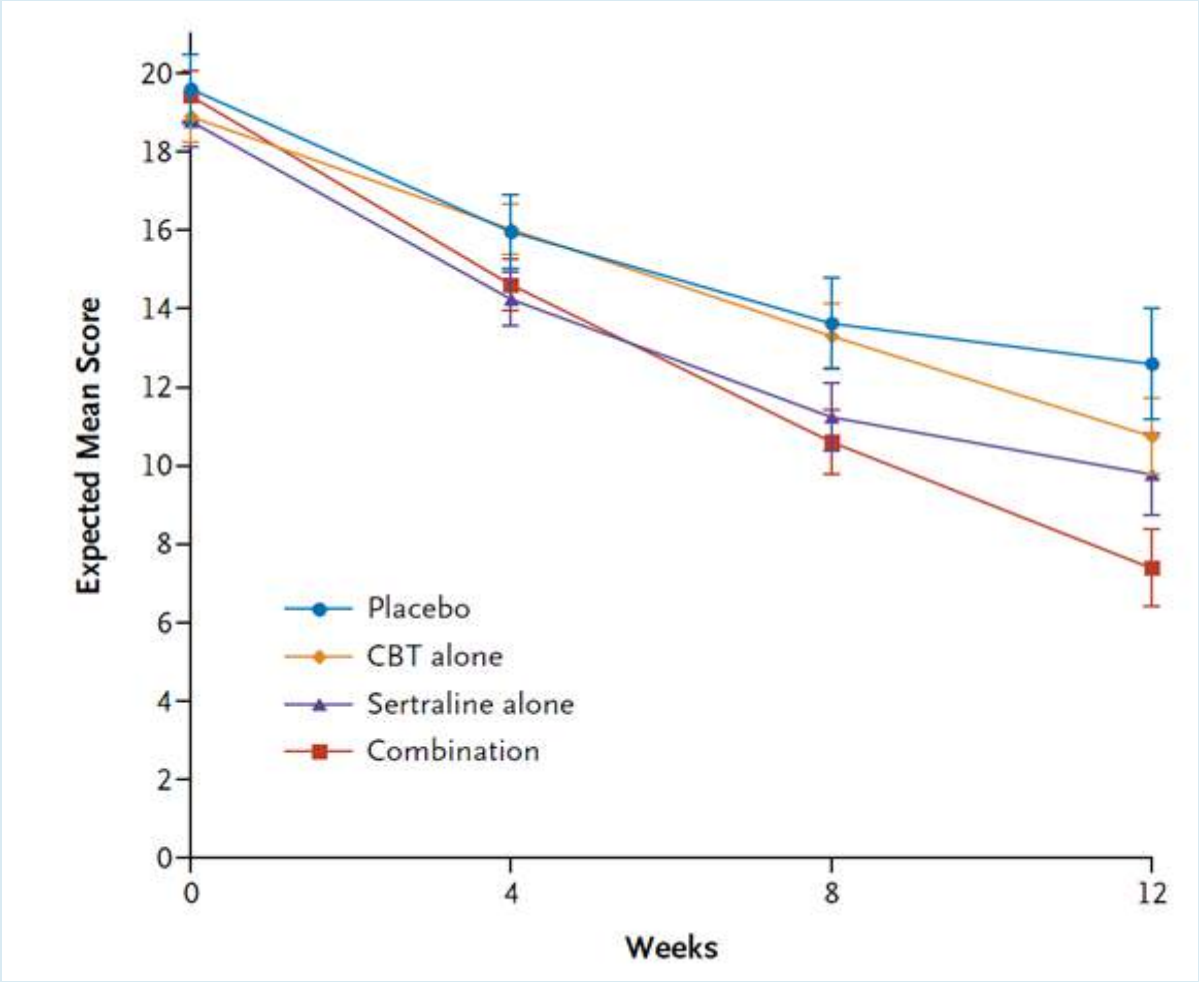
The odds ratio compares the odds of an event after exposure to some occurrence with the odds of that event in a control group not exposed to the occurrence. Often the exposure for the purposes of this chapter is the treatment being studied, whereas the control group is not exposed to the treatment. Consider a fictitious example: 25 adolescents with conduct problems receive a wilderness experience and a similar group of 20 adolescents with conduct problems receive outpatient counseling. One year later 3 adolescents in the wilderness group have been arrested versus 7 in the group receiving outpatient counseling. The odds ratio for not being arrested (the desired outcome) for comparing results in the wilderness experience to outpatient treatment is 22/3 divided by 13/7. This comes out to 7.33/1.86 = 3.94. In plain English, the teens exposed to the wilderness program were about four times less likely to be arrested during the year following the wilderness program.

USING YOUR UNDERSTANDING OF THESE BASIC CONCEPTS TO READ A PAPER ON A SINGLE RCT EFFICIENTLY

Ask yourself: *what is the population studied* and *what is the primary outcome as displayed in tables and visual graphics.* This is a highly efficient way of reading new studies. For example, consider the results from a study examining the response to treatment of children and adolescents with an anxiety disorder. Called the Child/Adolescent Anxiety Multimodal Study (CAMS; Walkup et al, 2008), it recruited children and adolescents from multiple sites with common anxiety disorders—generalized anxiety, social phobia, separation anxiety disorder—and randomized these youth into receiving one of four treatment strategies: CBT alone, sertraline alone, combined CBT and sertraline, and placebo. The primary outcome was the child's score on the Pediatric Anxiety Rating Scale (PARS).

Examine the graph in Figure A.6.8 showing PARS scores for the 4 groups, each receiving a different treatment. The red line descends consistently and rapidly with participants in the group treated with sertraline and CBT enjoying ongoing relief from anxiety symptoms (scores greater than 13 are consistent with an anxiety disorder). Combining sertraline and CBT, based on these data, is the most effective treatment. The *error bars* on the PARS figures show how much *scatter* there is in those numbers. The top of an error bar is most easily understood as the mean plus the standard error of the mean (calculated as the standard deviation divided by the square root of the number of participants), and the bottom of the error bar is the mean minus the standard error of the mean. More complex methods of calculating standard error represented by the error bars also exist. Note that the confidence interval bars for combination therapy (sertraline + CBT) do not overlap with any of the three other groups.

Figure A.6.8 Results from the Child/Adolescent Anxiety Multimodal Study, a trial randomizing anxious youth into one of 4 different treatments*



*Walkup et al, 2008; with permission.

In this case, the clinical magnitude of the impact of treatment on outcome was evaluated by calculating effect sizes for each of the three intervention groups compared to the placebo group: how different were each of the three intervention groups on the PARS compared to the placebo group? The effect size at 12 weeks was 0.86 for combination therapy, 0.45 for sertraline alone and 0.31 for CBT alone. The authors also calculated the number needed to treat. Recall that to calculate an NNT, we must choose a threshold or 'hurdle' for each subject to 'clear'. Higher hurdles are more difficult to clear and therefore will generate higher NNTs; lower hurdles will generate lower NNTs. The authors chose as a hurdle that the youth must be rated as either improved or very much improved on the PARS by the end of treatment, at 12 weeks. Using this criterion, the authors calculated the NNT for sertraline alone as 3.2; for cognitive behavioral therapy alone as 2.8; for the combination of both treatments the NNT was 1.7. This low NNT is important for clinicians. The authors also used statistics to convey to readers how certain they were that their results were not a lucky fluke caused by random events that together resulted in improvements in those youth receiving sertraline and CBT. To assess this possibility, the authors calculated the confidence intervals for each effect size

and for each number needed to treat. The confidence intervals for a 95% certainty around the calculated effect size of 0.86 were 0.56 to 1.15 for youth receiving the combination of sertraline and CBT compared to youth receiving placebo, a large effect. According to convention, effect sizes of 0.8 and higher indicate a large effect, 0.3 and below a small effect.

In summary, to read a paper on an RCT efficiently, look first at the population studied, then go right to the primary outcome results presented either graphically or in a table. Often the answer to the central research question posed by the study will be right there visually in the graph or as numbers in a table.

MAKING SENSE OF DATA SUMMARIES

Aggregated evidence summarizes data from multiple RCTs into a single data set and thereby provides a 'birds-eye view' of evidence to reach conclusions about what works for a specific condition or group of conditions. Importantly, the results may become more precise because the number of subjects increases. There are at least four different ways to aggregate data meaningfully: meta-analysis, systematic review, network meta-analysis, and distillation and matching. Let's consider each.

Meta-Analysis

A meta-analysis combines the results of multiple randomized trials, each using a specific intervention, to produce overall—or 'average'—statistics on the effect of those interventions regards benefits and harms, but there are many caveats. For example, if the populations studied are too different, the outcomes measured not sufficiently similar, or there is excessive bias in some studies, then the overall statistics may not be meaningful. A meta-analysis may look at how a class of treatments works for a class of disorders—how cognitive-behavioral treatment works for anxiety disorders in children, for example, or how antidepressants work for anxiety disorders. Crucially, 'forest plots' are graphic summaries of the results from a meta-analysis and are the best way to rapidly grasp results. For example, Figure A.6.9 shows a forest plot providing a splendid view over the effects of individual medications for ADHD compared to placebo (Cortese et al, 2018) summarizing:

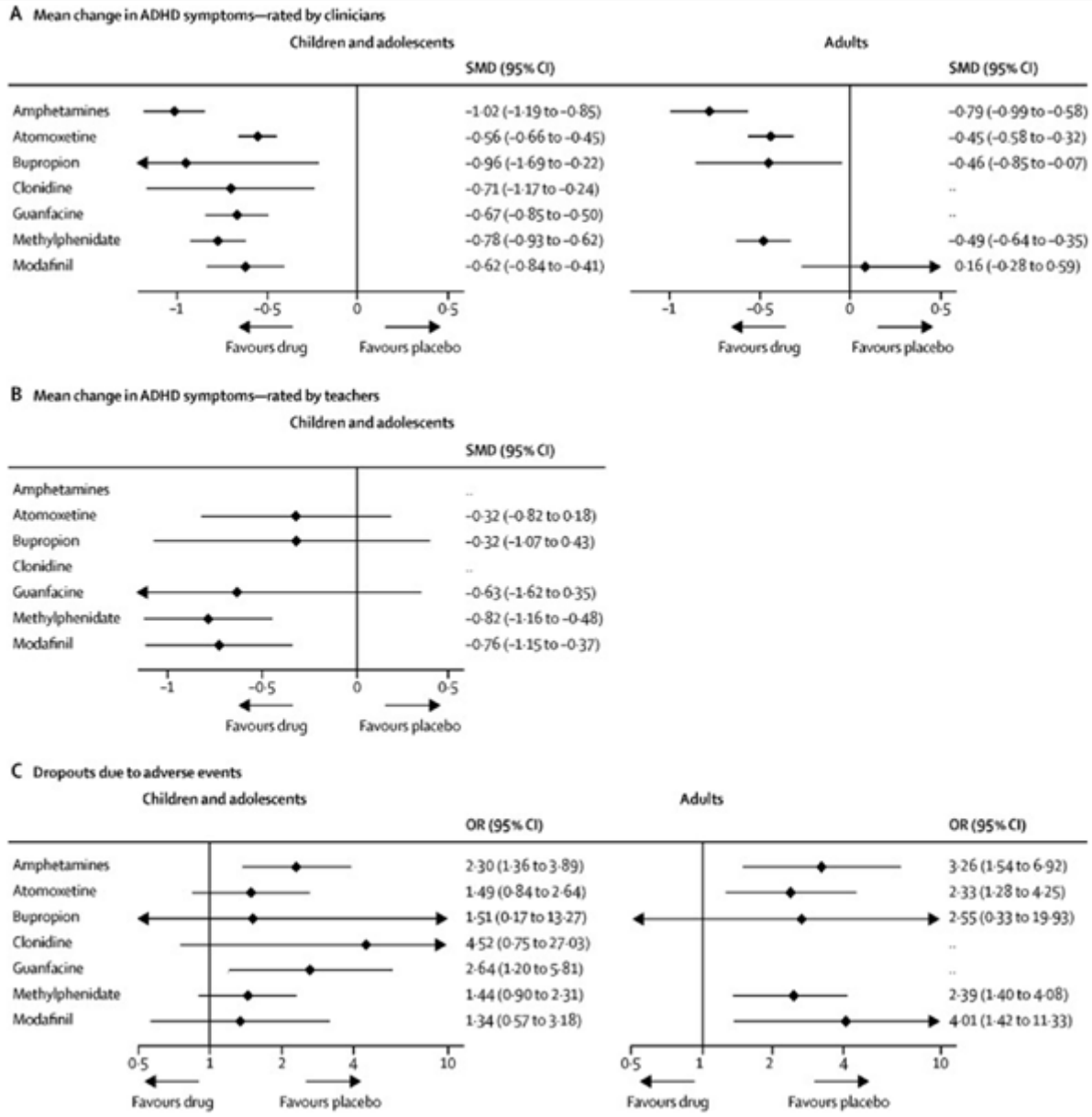
- Effects on ADHD symptoms as rated by clinicians (broken out separately for youth and adults),
- Effects on ADHD symptoms rated by teachers, and
- The chances of the family stopping medication compared to placebo.

This figure was created using a specific form of meta-analysis called *network meta-analysis*. The amount of useful information available at a glance is stunning. We will be discussing how to read a forest plot soon when dealing with systematic reviews.

Systematic Review

A systematic review is a meta-analysis completed using a well-defined methodology for the search, inclusion, and synthesis of studies. The statistical

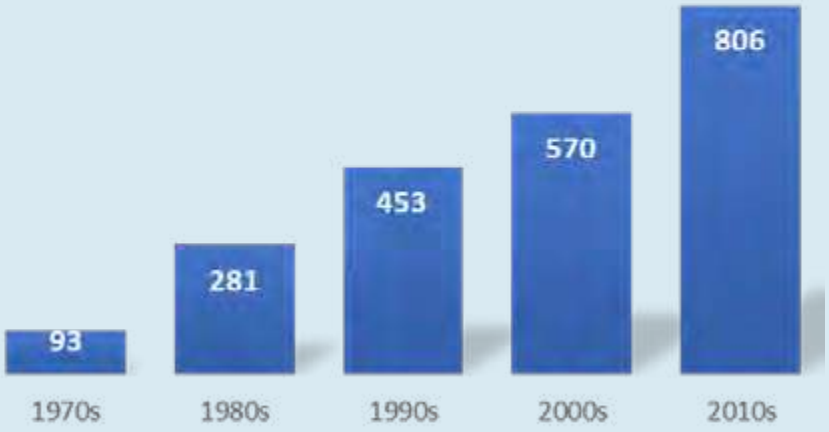
Figure A.6.9 Comparative efficacy and tolerability of medications for ADHD*



*Cohrane Database of Systematic Reviews (Cortese et al, 2018; with permission)

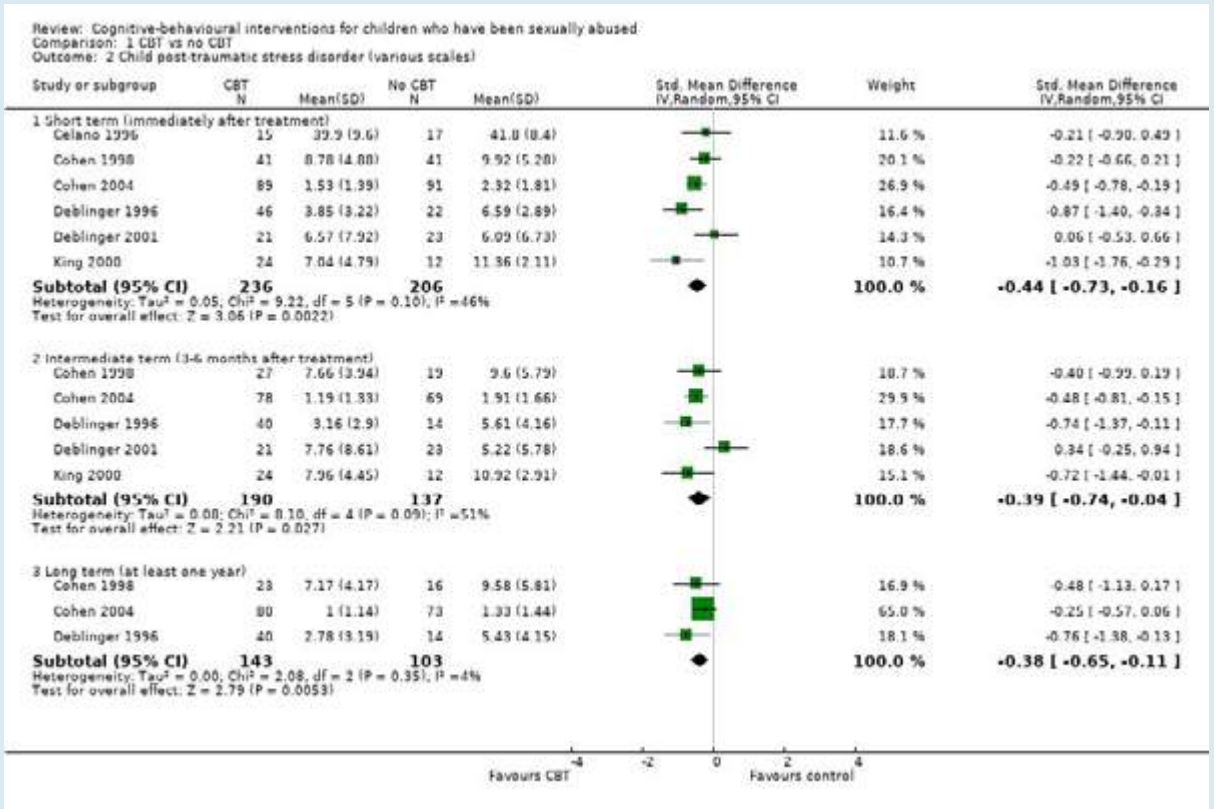
summary remains a meta-analysis but the reader has a quality guarantee about how it was created: the authors actually explain how the sausage was made. An analogy of the difference between a meta-analysis versus a systematic review is that the former is like a sausage-maker working with more freedom to make it his way, whereas the latter is making sausages in a glass kitchen with a detailed and published recipe. These are efforts to minimize potential bias. A systematic review done by the Cochrane Library, for example, follows a detailed process outlined in their training manual for reviewers and published on the Internet. Systematic reviews are increasingly popular in child psychiatry. Figure A.6.10 shows the increase in systematic reviews published in a leading child psychiatry journal by decade from the 1970s through the 2010s. Our field increasingly relies on systematic reviews to aggregate information into useful knowledge.

Figure A.6.10 Number of hits to the term 'systematic review' in the *Journal of the American Academy of Child & Adolescent Psychiatry* according to decade, from 1970 to 2020*



*Hamilton J, unpublished data.

Figure A.6.11 A systematic review of cognitive-behavioral interventions for children who have been sexually abused*



*Macdonald et al, 2012; with permission.

The number of systematic reviews published by the Cochrane Collaboration concerning the mental health of children has also increased markedly in the last decade. From 2000 through 2009 there were 26 systematic reviews which show up searching in the Cochrane Database of Systematic Reviews using the search terms *child* AND (ADHD OR Anxiety OR Depression OR PTSD OR Oppositional)*. This is a broad search and, in visually scanning the abstracts, not all 'hits' appear potentially useful, but many do. Remarkably, from 2010 through 2019 the same search terms retrieve 235 systematic reviews, an increase of 900 percent! There is currently much useful aggregated data available to us clinicians caring for children. Figure A.6.11 shows a recent systematic review of the results of using CBT for sexually abused children with its results presented as a forest plot (Macdonald et al, 2012). Let's discuss how to use it.

The most efficient way to examine this forest plot is to go straight to the 'black diamonds,' which summarize the overall effect. The first black diamond under the heading 'short term' (immediately after treatment) has a width, representing 95% confidence intervals, which does not overlap the '0' vertical line representing no effect, and the authors provide the standardized mean difference for the summary of these 6 studies as -0.44. This number (0.44), is called the standardized mean difference, or effect size, or Cohen's *d*—these are all the same thing. The treatment group improved on average in the various measures used to measure PTSD symptoms 0.44 of a standard deviation, compared to the control group. This 0.44 effect size is generally considered as moderate. Similar conclusions can be quickly drawn from this plot both for intermediate-term results and long-term results, demonstrating how efficiently a forest plot displays information.

Therefore, when looking at a systematic review, consider going straight to the figures and have a look. These graphic displays of information are easily digestible and cut to the core results of the review immediately. With journals publishing increasing numbers of systematic reviews, it becomes easier to find such succinct summaries. There is one caveat: because systematic reviews are labor intensive, it is often necessary to do the work of updating their results by searching for RCTs from more recent years on your own.

Network Meta-Analysis

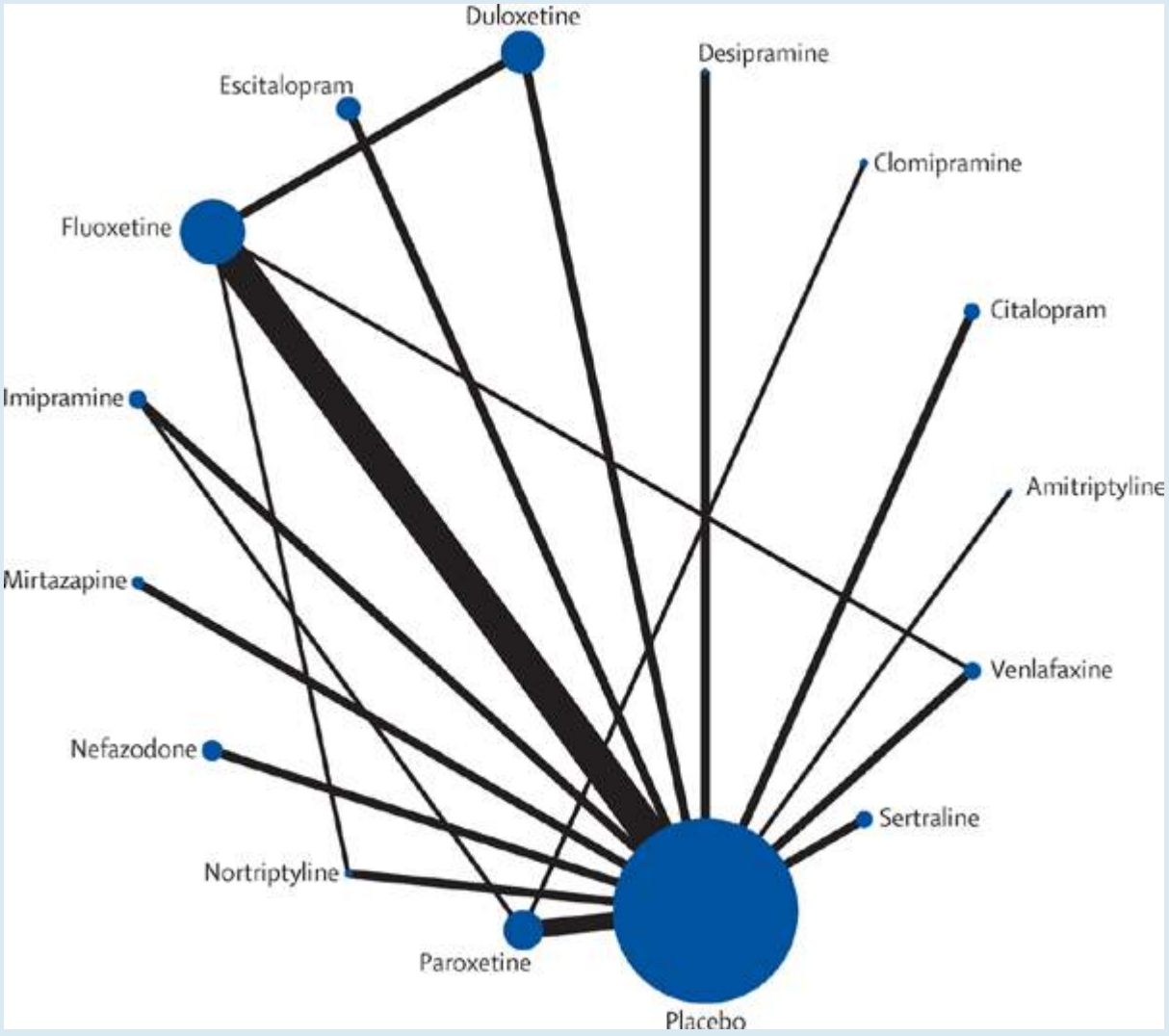
Another method of aggregating data is called network meta-analysis (NMA), widely used to summarize data about response to treatment. Network meta-analyses, unlike meta-analyses or systematic reviews, compare the effectiveness of multiple treatments using methods which allow rank-ordering of those treatments from most effective to least effective. For clinicians choosing a treatment, this is a significant improvement. Unlike a meta-analysis, which evaluates the effectiveness of a single treatment measured against a control, a network meta-analysis can evaluate the effectiveness of multiple treatments *measured against each other* in a single analysis. A network meta-analysis does this by combining what is called 'direct evidence'—evidence obtained from head-to-head studies comparing one active treatment to another active treatment—with 'indirect evidence'—evidence inferred via computation about the effectiveness of a treatment (A) relative to another (B) because both, for example, have been compared to a third treatment (C), where 'C' may be either a control or an active intervention. The multiple

For a discussion of network meta-analysis as used by Cochrane go here. If you are new to Cochrane, you will need to register here.

studies of varied treatments for a similar condition, tested either against each other as an active treatment or a placebo or waiting list, comprise a network of information about effectiveness (Rouse et al, 2017).

Consider how results of a network meta-analysis of antidepressants studied in adolescents with major depressive disorder can help answer the question: which antidepressant is most likely to be helpful for Alex, the depressed adolescent who tried to hang himself mentioned earlier. Going to PubMed and typing *network meta-analysis adolescen* major depression* (where * is a 'wild card'), the first article listed contains Figure A.6.12 (Cipriani et al, 2016) whether to use pharmacological interventions in this population and which drug should be preferred are still matters of controversy. Consequently, we aimed to compare and rank antidepressants and placebo for major depressive disorder in young people. We did a network meta-analysis to identify both direct and indirect evidence from relevant trials. We searched PubMed, the Cochrane Library, Web of Science, Embase, CINAHL,

Figure A.6.12 A graphic representation of effectiveness of medications used to treat major depressive disorder in adolescents, together making up a 'network' to be used in a network meta-analysis*



*Cipriani et al, 2016; with permission.

PsycINFO, LiLACS, regulatory agencies' websites, and international registers for published and unpublished, double-blind randomised controlled trials up to May 31, 2015, for the acute treatment of major depressive disorder in children and adolescents. We included trials of amitriptyline, citalopram, clomipramine, desipramine, duloxetine, escitalopram, fluoxetine, imipramine, mirtazapine, nefazodone, nortriptyline, paroxetine, sertraline, and venlafaxine. Trials recruiting participants with treatment-resistant depression, treatment duration of less than 4 weeks, or an overall sample size of less than ten patients were excluded. We extracted the relevant information from the published reports with a predefined data extraction sheet, and assessed the risk of bias with the Cochrane risk of bias tool. The primary outcomes were efficacy (change in depressive symptoms).

The *thickness* of each connecting line in Figure A.6.12 reflects the *number of studies;* the size of each node reflects the *total number of participants* assigned to that condition. This network meta-analysis compares multiple treatments (in this case different antidepressants) by using both direct comparisons of interventions within randomized controlled trials (i.e., head-to-head studies between antidepressants) *and* indirect comparisons across trials based on a common comparator. The term 'network' refers to the whole system of multiple studies comparing results for multiple treatments. A network meta-analysis then uses statistical processes to examine both direct and indirect comparisons between each pair in the network (Nikolapopoulou et al, 2018). For example, in this network of RCTs studying antidepressants in adolescents, there is no direct study comparing sertraline to fluoxetine, yet the network meta-analysis will use statistical processes to generate a comparison between the two based on the results when each was compared to placebo. The analysis then can rank order each treatment based on effectiveness as well as tolerability (e.g., based on the proportion of subjects stopping the treatment) using a statistical approach called Surface Under the Cumulative Ranking curve (SUCRA)—highly useful information for clinical practice.

This network meta-analysis generates results useful in considering how to treat Alex because it rank-orders the antidepressants for efficacy as well as for tolerability and effect on suicidal ideas/behavior relative both to each other and to placebo. Network meta-analyses can rank-order treatments in a way that conventional pairwise meta-analysis cannot, as the technique borrows strength from indirect evidence to gain certainty about all treatment comparisons and allows for estimation of comparative effects that have never been investigated head-to-head in randomized clinical trials (Mills et al, 2013). See Figure A.6.13 for a more detailed explanation of network meta-analysis.

SUCRA

The statistical approach used to rank-order a treatment for either effectiveness or tolerability, called Surface Under the Cumulative Ranking curve, often abbreviated as SUCRA, uses a graphic representation similar to the receiver operating characteristic (ROC) analysis used in diagnosis (click here to view a tutorial on ROC analysis). The statistical program considers possible ways to rank-order these 14 treatments and asks how closely each possible ranking matches the data from the available studies. The 'area under the curve' is the metric used to measure how well any given ranking matches the available data.

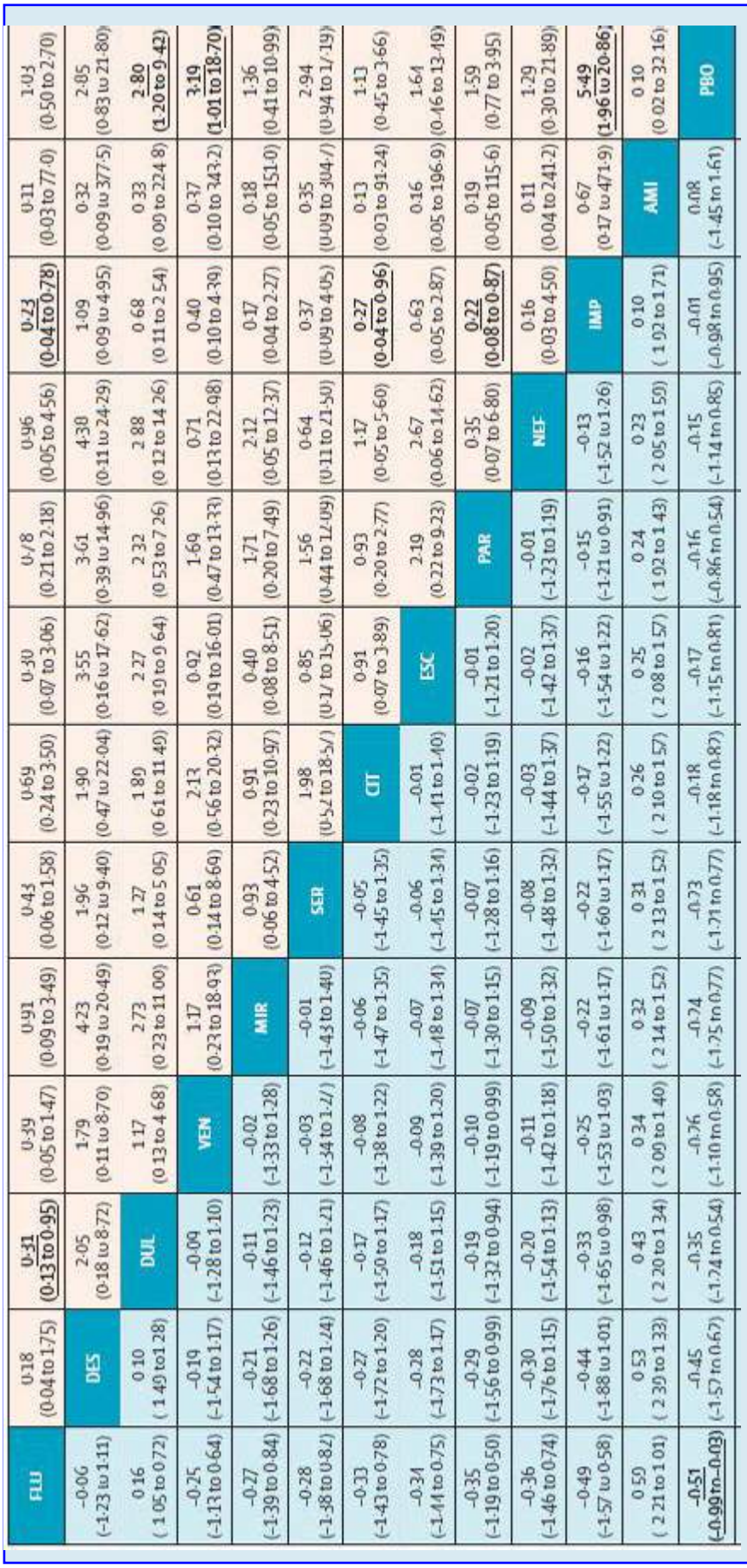
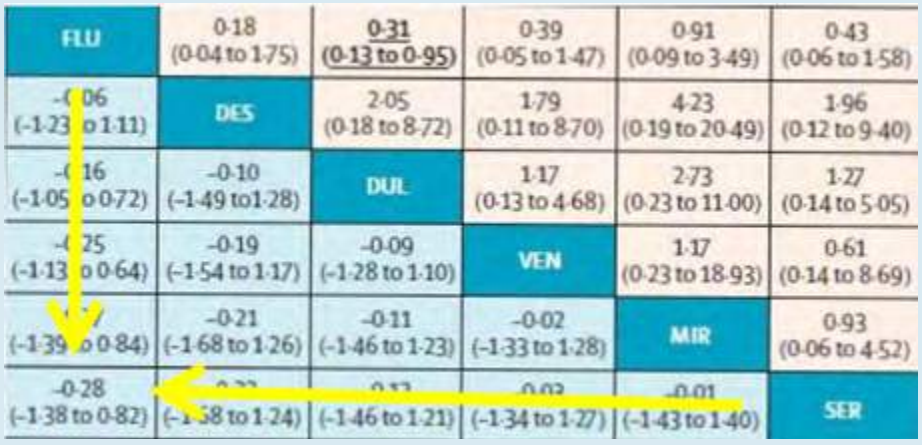


Figure A.6.13 Network meta-analysis of eficacy and tolerability*

This table shows the results of a network meta-analysis that compares 14 different drugs and placebo when treating depression in youth, regards both treatment response (in blue) and tolerability (pink). Only one drug—fluoxetine—reaches statistical significance for reducing symptoms, with a standard mean difference of -0.53 and confidence intervals of -0.99 to -0.03. Tolerability is measured by the odds ratio of discontinuing treatment. An odds ratio below '1' means the row defining treatment was tolerated better than the column defining treatment. Efficacy results (in light blue) are based on mean overall change in depressive symptoms and have been standardized so that results from different studies using different scales can be compared with one another.

Drugs are reported in order of eficacy according to the Surface Under the Cumulative Ranking curve, or SUCRA (see below). Comparisons are read from left to right. The efficacy (light blue boxes) and tolerability (pink boxes) estimates are located at the intersection of the column-defining treatment and the row-defining treatment, expressed as the standardized mean difference (with confidence intervals in parentheses). For eficacy (mean overall change in symptoms), a standardized mean difference below '0' favors the column-defining treatment. For tolerability (discontinuation due to adverse events), an odds ratio below '1' favors the row-defining treatment. Significant results are in **bold** and underlined. AMI=amitriptyline; CIT=citalopram; DES=desipramine; DUL=duloxetine; ESC=escitalopram; FLU=fluoxetine; IMP=imipramine; MIR=mirtazapine; NEF=nefazodone; OR=odds ratio; PAR=paroxetine; PBO=placebo; SER=sertraline; SMD=standardized mean diference; SUCRA=surface under the cumulative ranking curve; VEN=venlafaxine. Continues next page.

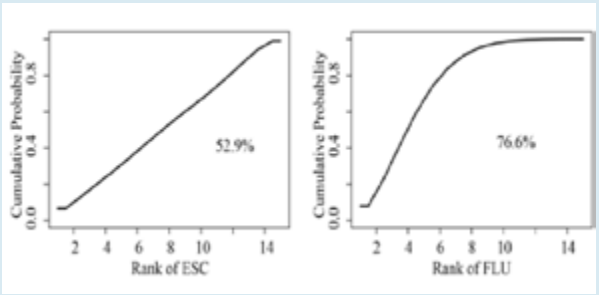
*Cipriani et al, 2016; with permission. Note that three drugs have been omitted.



The dark blue blocks in Figure A.6.13 contain the treatments being compared. The diagonal 'descending staircase' in dark blue is very useful for clinicians (read left to right). It rank-orders the treatments based on SUCRA data for efficacy, from most to least efficacious. The light blue blocks contain statistically generated estimates of the standardized mean differences in the change in depression ratings comparing the effectiveness of the medication in the column with that on the row. For example, in this fragment from Figure A.6.13, at the intersection of the column for fluoxetine (FLU) and the row for sertraline (SER) the standardized mean difference is -0.28 (that is, below 0—favoring fluoxetine). The confidence intervals for this standardized mean difference are -1.38 to -0.82 and contain zero, which would represent no statistically significant difference. That is, according to this network meta-analysis, fluoxetine is more effective than sertraline but the difference is not statistically significant. The standardized mean difference was used because depression rating scale data are continuous rather than categorical. The pink blocks represent data the authors call 'tolerability' because the figures are based on the proportion of youth who continue to take the drug rather than discontinue it. Since continuing vs discontinuing is a categorical variable, the results are presented as an odds ratio. In this analysis you can see also that fluoxetine (FLU) is less likely to be discontinued than duloxetine (DUL) (OR = 0.31) and the difference is statistically significant (95% confidence intervals do not contain 1).

Figure A.6.14 shows the SUCRA curves for fluoxetine and escitalopram. The x-axis contains all the competing medications and placebo. A rising curve shows that the medication being assessed is winning its comparisons and therefore generates more 'area under the curve.' Note how quickly the fluoxetine SUCRA curve rises compared to other drugs, whereas escitalopram seems to plod along. This view is too simplistic, of course, since the actual curves are generated via complex statistical methods. The visual information is nevertheless helpful in grasping a sense of how medications compare to each other for effectiveness. A SUCRA value of 50% (or 0.50) represents a treatment that has as many wins as losses in comparisons, and a value 100% (or 1.0) is a perfect score for a superior treatment winning all its comparisons.

Figure A.6.14 Surface Under the Cumulative Ranking (SUCRA) data for comparing effectiveness of 14 different antidepressants. Results for escitalopram (ESC) and fluoxetine (FLU)*



*Cipriani et al, 2016; with permission.

A SUCRA score for each drug can be interpreted as the proportion of treatments worse than the one of interest. Figure A.6.15 shows SUCRA scores for safety from suicidal behavior or ideation in children for a range of antidepressants.

Thanks to network analysis, we now can rank-order antidepressants in youth for efficacy, for risk of increasing suicidal ideas or behavior, and for tolerability—each one helpful in choosing a medication for Alex. These data suggest that

fluoxetine ranks first for efficacy, first for tolerability, and towards the middle of the pack for safety from suicidal ideas or behavior. In addition, duloxetine deserves consideration since it ranks high for efficacy and does well regards safety. Alex was started on fluoxetine with duloxetine as a potential back-up if fluoxetine failed.

Figure A.6.15 Safety from suicidal behavior or ideation in children for a range of antidepressants. Larger Surface Under the Cumulative Ranking curve (SUCRA) scores reflect increased safety*

DRUG	SUCRA Score (%)
Imipramine	68.9
Placebo	65.6
Duloxetine	65.3
Escitalopram	60.4
Clomipramine	59.7
Fluoxetine	53.3
Paroxetine	49.0
Citalopram	42.7
Sertraline	28.0
Venlafaxine	2.8

*Data from Cipriani et al, 2016

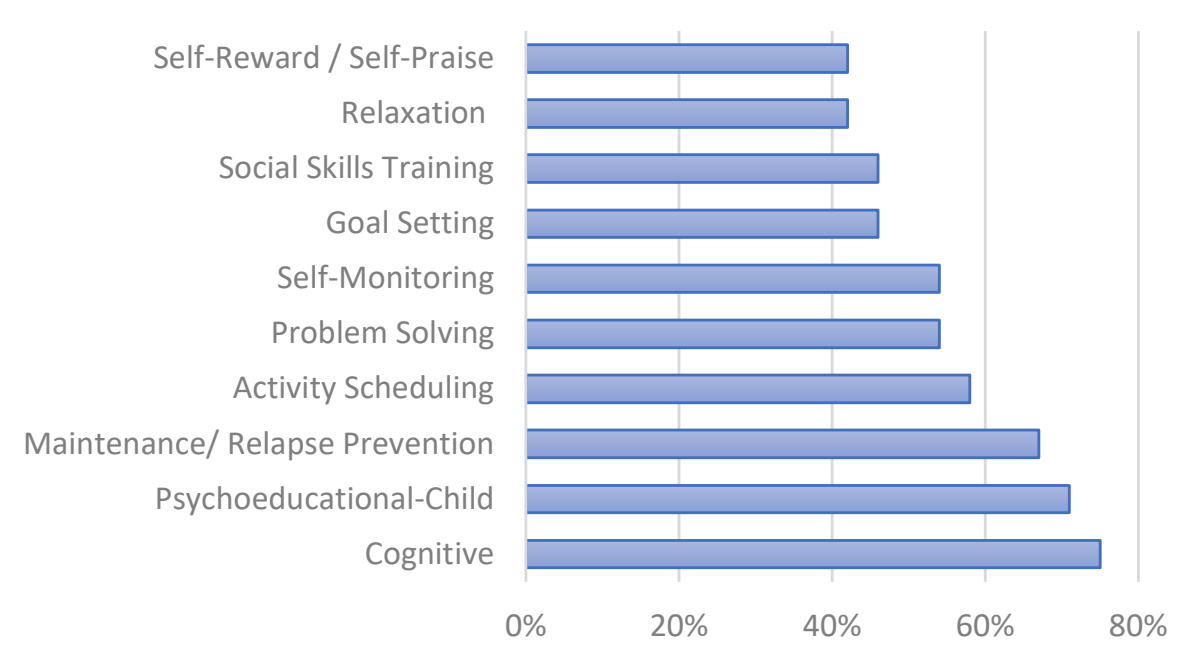
Distillation and Matching

A fourth method of aggregating data—appropriate only for studies of psychotherapy (not medicines)—in order to summarize what works for a specific disorder is to 'distill' treatment protocols (manuals) which have done well in RCTs for a specific disorder. 'Distilling' here refers to breaking up often large psychotherapy manuals into succinct specific elements like 'relaxation techniques' or 'restructuring cognitions' or 'encouraging pleasurable activities.' In other words, rather than choosing a specific manual based on one or more studies, why not lift the hood of all the manuals proven to get a good result and look at what's going on in them: what are the aspects that make those manuals work? That is exactly what Chorpita and Daleiden did: they chose to 'aggregate evidence-based treatment protocols empirically according to their constituent treatment procedures' (Chorpita et al, 2005). That is, they looked for common elements or procedures that showed up in multiple manuals that were effective in treatment trials. They reasoned that such elements occurring in multiple effective manuals were likely to be a part of what created positive change. If you were studying a fleet of race cars to build a great race car, you might examine multiple cars which were winning races and then look for features in their engines which they shared (common elements).

Chorpita and Daleiden began by saying that a psychotherapy manual was 'winning' if, in at least one study, '…an active, nonpharmacological treatment beat one or more other study groups (e.g., psychosocial treatment, medication, combined psychosocial and medication, placebo, waiting-list, no treatment, or other control group) in a randomized trial on the primary outcome measure in the target symptom domain (e.g., the primary depression measure in a study of depression)' (Chorpita et al, 2005). For example, if most manuals proven to help with depression have a plan for how the therapist will intervene by challenging pessimistic, gloomy thoughts as unrealistic, this is called a common element (cognitive, because the goal is to gently challenge unrealistic, pessimistic cognitions).

The results of their distilling (i.e., summarizing) the elements from 'winner' manuals for treating depressed youth is summarized in Figure A.6.16. Each bar represents the frequency, as a percentage, of all 'winning' manuals for depressed youth in which those specific elements appear. For example, 74% of the 'winning' manuals for depressed youth use cognitive therapy as one element of their protocol. These results 'match' as potentially effective treatments for individual youth with depression but can be further tailored based on the individual's gender, age, and race to improve the match (Chorpita & Daleiden, 2009). Chorpita and Weisz, (2009) went through a similar process to define the treatment elements most frequently appearing in 'winner' manuals for youth with anxiety, with oppositional defiant behavior, and for those with trauma. For example, here are five elements

Figure A.6.16 Top 10 treatment elements ranked by how often (%) each appears in the manuals of 'winner' treatments for depressed youth*



*Based on data from Chorpita & Daleiden, 2009.

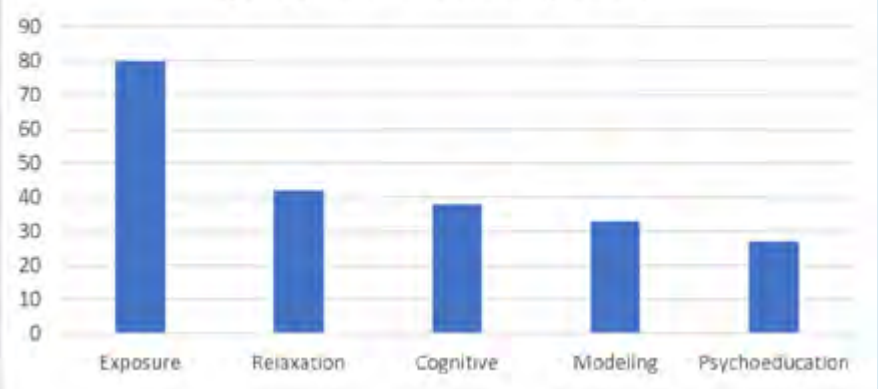
common in successful treatment manuals for anxious youth: exposure, relaxation, cognitive approaches, modeling, and psychoeducation about the nature of anxiety problems.

A strength of the distillation and matching approach is its flexibility in the treatment of youth with comorbid symptoms. A protocol for how to deliver treatment elements for anxiety, depression, trauma, and conduct problems, including youth with comorbid symptoms, allows for individualizing treatment within the context of elements proven to appear frequently in 'winner' manuals (Chorpita & Weisz, 2009). To understand how this works in an individual case, recall the image mentioned earlier of an old-fashioned liquor still, boiling down a big mash of grain to produce a small bottle of liquor. Similarly, here, multiple large bulky psychotherapy manuals, chosen because they are proven to be effective for a particular problem, are boiled down to their essential ingredients—what the therapist is directed to do. Figure A.6.17, for example, shows that 80% of manuals effective for anxiety disorders use exposure. Exposure is therefore a common element of most psychotherapy manuals for anxiety

Distillation and Matching Treatment Elements for Depression as a Feasible Evidence-Informed Treatment for Alex

It's a long way from the cool elegance of a Cochrane systematic review to the reality of a child's life, from results in a meta-analysis to what is feasible in our offices. The messiness of real life and the context of our workplace often shape what is feasible. Recall Alex, the high school student who tried to hang himself and who remains depressed. He hated his one-week stay in the hospital and does

Figure A.6.17 Frequency (%) of common elements effective in treating anxious youth distilled from 'winner' manuals*



*Chorpita & Daleiden, 2009.

Evidence-informed rather than evidence-based allows us to be more realistic since a therapy is useless if it is not feasible.

not want to miss any more school. He lives with his mother while his older brother is away in college. He is estranged from his father, an electrical engineer. His Beck Depression Inventory score is 30, consistent with severe depression. A meta-analysis suggests mentalization as a potentially useful psychotherapy to reduce his risk of self-harm. A network meta-analysis offers evidence that fluoxetine is the best antidepressant for him. What evidence-informed psychotherapy is feasible in our setting to combine with fluoxetine to improve his mood? Note we are hedging here away from 'evidence-based' to 'evidence-informed' and adding the modifier 'feasible'. Evidence-informed rather than evidence-based allows us to be more realistic since a therapy is useless if it is not feasible.

The therapist treating Alex used the distillation and matching approach described above. For Alex's severe depression, distillation and matching suggests using components or elements of treatment which appear often in 'winner' manuals shown in at least one RCT to be effective in treating depression. The ten most prevalent elements in successful manuals treating depressed youth are presented in Figure A.6.16, namely, cognitive, psychoeducation, maintenance and relapse prevention, activity scheduling, problem-solving, self-monitoring, goal-setting, social skills training, relaxation exercises, self-reward and self-praise.

Many of these treatment elements for depressed youth can be introduced into early interviews with Alex, split between Alex alone and Alex with his mother. *Psychoeducation* in this context refers to teaching them both key aspects of what is known about severe depression: that it often affects physiological functions like appetite, sleep, the ability to feel pleasure and joy, concentration, energy level, motivation, and decision-making among others; that it is normal to lose interest during depressive episodes, or be unable to concentrate on studies, or have the energy to complete projects. Encouraging Alex to gently challenge more pessimistic and less hopeful thoughts is a *cognitive* intervention and encouraging him to be a good observer of his thoughts and moods is *self-monitoring*. *Activity scheduling* include encouraging him to find time to hang out with friends to talk, a hedonic activity for him. Alex had a complete recovery in his functioning

over the course of a year-long psychotherapy based on the common elements for depression and treatment with fluoxetine.

Click on the picture to view a video tutorial on how to search using PubMed (2:27)

The Cochrane Database of Systematic Reviews (CDSR) and PubMedCentral (PMC) are free, easily available allies in finding evidence

Feasibility begins with finding convincing evidence quickly using databases in the public domain. Two Cochrane databases—the Cochrane Database of Systematic Reviews (CDSR) and the Cochrane Central Register of Controlled Trials (CENTRAL) are useful, so is PubMedCentral and PsycINFO.

All these databases provide abstracts for free, but *PubMedCentral has the advantage of containing not only abstracts but also full-text access to literally millions of journal articles.* CDSR offers abstracts and 'plain language' summaries of the conclusions of each review at no charge worldwide. In low-income countries, Cochrane offers full text versions of both CENTRAL and CDSR at no charge. The exhaustive thoroughness of Cochrane's systematic reviews gives them high credibility. PsycINFO contains free abstracts from a wide swath of journals in psychology.

REFERENCES

Andrade C (2015). Understanding relative risk, odds ratio, and related terms: as simple as it can get. *The Journal of Clinical Psychiatry* 76: e857–61. https://doi.org/10.4088/JCP.15f10150

Chorpita BF, Daleiden E, Weisz J (2005). Identifying and selecting the common elements of evidence-based interventions: a distillation and matching model. *Mental Health Services Research* 7: 5–20

Chorpita BF, Daleiden EL (2009). Mapping evidence-based treatments for children and adolescents: application of the distillation and matching model to 615 treatments from 322 randomized trials. *Journal of Consulting and Clinical Psychology* 77:566–579. https://doi.org/10.1037/a0014565

Chorpita BF, Weisz JR (2009). *MATCH-ADTC: Modular Approach to Therapy for Children with Anxiety, Depression, Trauma, or Conduct Problems*. Satellite Beach, FL: PracticeWise, LLC

Cipriani A, Zhou X, Del Giovane C et al (2016). Comparative efficacy and tolerability of antidepressants for major depressive disorder in children and adolescents: a network meta-analysis. *The Lancet* 388 (10047): 881–890. https://doi.org/10.1016/S0140-6736(16)30385-3

Cortese S, Adamo N, Del Giovane C et al (2018). Comparative efficacy and tolerability of medications for attention-deficit hyperactivity disorder in children, adolescents, and adults: a systematic review and network meta-analysis. *The Lancet Psychiatry* 5:727–738. https://doi.org/10.1016/S2215-0366(18)30269-4

Guyatt GH, Rennie D (2002). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice* Chicago: AMA Press

James AC, James G, Cowdrey FA et al (2015). Cognitive behavioural therapy for anxiety disorders in children and adolescents. *Cochrane Database of Systematic Reviews.* https://doi.org/10.1002/14651858.CD004690.pub4

Kabai P (2021). *Bias and Random Error* (blog). http://apollo.lsc.vsc.edu/classes/remote/lecture_notes/measurements/bias_random_errors.html

Macdonald G, Higgins JPT et al (2012). Cognitive-behavioural interventions for children who have been sexually abused. *The Cochrane Database of Systematic Reviews.* https://doi.org/10.1002/14651858.CD001930.pub3

March J, Silva S, Petrycki S et al (2004). Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: Treatment for Adolescents With Depression Study (TADS) randomized controlled trial. *JAMA* 292:807-820. doi: 10.1001/jama.292.7.807. PMID: 15315995

Mills EJ, Thorlund K, Ioannidis JPA (2013). Demystifying trial networks and network meta-analysis. *BMJ* 346 (may14 2): f2914–f2914. https://doi.org/10.1136/bmj.f2914

Nikolakopoulou A, Mavridis D, Furukawa TA et al (2018). Living network meta-analysis compared with pairwise meta-analysis in comparative effectiveness research: empirical study. *BMJ*, February, k585. https://doi.org/10.1136/bmj.k585

Ougrin D, Tranah T, Stahl D et al (2015). Therapeutic interventions for suicide attempts and self-harm in adolescents: systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry* 54: 97-107.e2. https://doi.org/10.1016/j.jaac.2014.10.009

Rouse B, Chaimani A, Li T (2017). Network meta-analysis: an introduction for clinicians. *Internal and Emergency Medicine* 12: 103–11. https://doi.org/10.1007/s11739-016-1583-7

Walkup JT, Albano AM, Piacentini J et al (2008). Cognitive behavioral therapy, sertraline, or a combination in childhood anxiety. *New England Journal of Medicine* 359:2753–2766. https://doi.org/10.1056/NEJMoa0804633