Generalizing from Clinical Trial Data: A Case Study. The Risk of Suicidality Among Pediatric Antidepressant Users

Running Title: *Generalizing from Clinical Trial Data: A Case Study*

Authors: Joel B. Greenhouse[1], Eloise E. Kaizar[2], Kelly Kelleher[3], Howard Seltman[1], William Gardner[3]

Affiliations:
[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA
[2] Department of Statistics, Ohio State University, Columbus, OH
[3] Center for Innovation in Pediatric Practice, Columbus Children's Research Institute, Columbus, OH

Grant Support: This study was funded in part by grant 1R01 MH7862 from the NIMH.

Keywords: External Validity; FDA; Meta-analysis; Randomized Controlled Clinical Trials; Youth Risk Behavior Survey

Corresponding Author:
Joel B. Greenhouse, Ph.D.
Department of Statistics
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
Voice: 412-268-8872
FAX: 412-268-7828
Email: joel@stat.cmu.edu

> [T]he carefully controlled clinical trials currently conducted premarket under the existing statutory framework consists of study populations that are commonly different in composition and health status from populations that will use the marketed drug. Study populations are chosen for a legitimate reason: to make data from the trials clearer and thus to make safety and efficacy testing more efficient. After approval, drugs are used by larger and more heterogeneous populations ...
>
> "The Future of Drug Safety: Promoting and Protecting the Health of the Public", *IOM Report*, 2007, pp. 153.

Summary

For the results of randomized controlled clinical trials (RCTs) and related meta-analyses to be useful in practice, they must be relevant to a definable group of patients in a particular clinical setting. To the extent this is so, we say that the trial is generalizable or externally valid. Although concern about the generalizability of the results of RCTs is often discussed, there are few examples of methods for assessing the generalizability of clinical trial data. In this paper, we describe and illustrate an approach for making what we call generalizability judgments and illustrate the approach in the context of a case study of the risk of suicidality among pediatric antidepressant users.

1. Introduction

Randomized clinical trials (RCTs) are the gold standard for generating evidence about the efficacy and effectiveness of interventions in medicine. Critical elements of trial design, however, are essential for making unbiased inferences about treatment effects obtained from those trials. One of these critical elements is the extent to which the sampling and recruitment strategies yield a sample population that is representative of the target population [1, 2]. Although concern about the generalizability of the results of RCTs is often discussed (e.g., [3]), there are few examples of methods for assessing the generalizability of clinical trial data or examples where subject selection effects make a difference. In this paper we illustrate both in the context of a case study of the risk of suicidality among pediatric antidepressant users. We begin with the background for the case study.

Case Study: Background

Newer antidepressants and related drugs are among the most widely prescribed and fastest growing classes of medication. During the period from 1998 to 2002, the prevalence of SSRI antidepressant use in children and adolescents nearly doubled from 0.93% to 1.66% [4]. Starting in 2005, however, this trend began to change after the Food and Drug Administration (FDA) issued a black box warning advising physicians and patients that antidepressants are associated with an increased risk of suicidal ideation and behavior in youth. Following the back box warning there has been a reported decrease in the number of prescriptions for antidepressants [5, 6] and a reported increase in suicidal events in

this age group [7], not to mention an increase in concern among clinicians, patients and families about the adverse consequences of losing treatment options.

The basis for the FDA's black box warning was a meta-analysis of 24 randomized placebo-controlled efficacy trials involving nearly 4600 children and adolescents, approximately half treated with SSRIs and other atypical antidepressants. The FDA analysis using study reports of adverse events found that youths receiving antidepressants had approximately twice the amount of suicidal ideations and behaviors than youths in the placebo control groups. There were no completed suicides.

In contrast, several recent observational studies have found a decrease in risk of suicide and suicidal behaviors among persons treated with antidepressants. Valuck et al. [8] using insurance claims and a filled prescriptions database followed over 24,000 adolescents newly diagnosed with a major depressive disorder (MDD) to investigate the relationship between antidepressant use and the risk of suicide attempt. They did not find a statistically significant increase in the risk of suicide attempts among antidepressant users even after adjusting for confounding variables. Simon et al. [9] used computer health plan records to identify 65,000 depressed patients including adolescents in a new index episode of antidepressant treatment. The outcomes of interest were suicide attempt with hospitalization and suicide death. They found that the risk of suicide attempt was highest in the month before starting antidepressant treatment and declined progressively after starting medication. Further, the risk of death by suicide was not significantly higher in the month after starting medication than in subsequent months. Using county-level vital statistics and county-level antidepressant prescription rate data, Gibbons et al. [10, 11] found that after adjusting for access to mental health care and demographic factors higher SSRI prescription rates were associated with lower suicide rates in children and adolescents.

Cause-and-effect relationships can not be established from observational studies such as these. Nevertheless, such studies are large, typically more representative of the population of interest and include more suicidal events than an RCT. Understanding why the results from these observational studies differ from the results of the FDA meta-analysis is a key question we consider in the remainder of this paper.

Methodological Overview

One possible explanantion for the difference in results between the observational studies and the FDA's meta-analysis could be due to the fact that these studies are sampling from different patient populations. For example, Kaizar et al. [12] have expressed concern about the generalizability of the findings from the FDA's meta-analysis due to the explicit exclusion of patients at high risk of suicide from most of the randomized trials, a critical subset of the population affected by the black box warning. Specifically, 14 of the 16 depression trials included in the FDA meta-analysis excluded youth with an elevated baseline risk for suicide [13]. To the degree that these exclusion criteria restricted the representativeness of the study samples, any generalizations to a larger target population could be limited.

A strategy for assessing the generalizability of samples obtained in randomized trials is to compare study samples to target populations on measured characteristics that might directly modify outcomes or are correlated with unmeasured variables that might modify outcomes. If the treatment interacts with some characteristic to produce an outcome that depends on the treatment-by-characteristic combination, a situation called treatment heterogeneity, then generalizing results from a study to the larger target population may not be appropriate [14]. Thus, if such a comparison shows that the study population is similar to the target population, then we would be more confident in making inferences from the former to the latter. If, however, the two populations are shown to be dissimilar and there is treatment heterogeneity there would be reason to be cautious about generalizing the study results.

Using this approach, Stevens et al. [15] studied the representativeness of the sample from the multi-site Multimodal Treatment of ADHD (MTA) trial. Their findings suggested that the trial sample was similar to community-based epidemiologic cohorts on demographic characteristics, although some differences in disease presentation were noted. This result may not be surprising since the investigators of the MTA intentionally sought to include as many sample sites as possible and to have as few exclusion criteria as they could.

In this paper, we describe and illustrate this approach for making what we call generalizability judgments. Specifically, we are concerned with the representativeness of RCT patients with respect to the target population of interest. Using the FDA meta-analysis of the risk of suicidality among pediatric antidepressant users as a case study, we examine the similarity of characteristics of the patients who participated in the trials used in the FDA meta-analysis to characteristics of the general population of youth with depression. We are interested in how well the samples from the randomized trials represent the target population of community youth with depression. We compare the characteristics of the youths in the trials with those of a nationally representative sample of youth completing the Youth Risk Behavior Survey (YRBS) who responded positively to a screening question for depression. We first consider group differences in the available demographic variables that may either directly or indirectly moderate the suicidality outcome. Second, we compare the rates of suicidality, since significant differences in the prevalence of suicidality may imply differences in underlying moderating variables between the two populations.

Our interest in this case study for making generalizability judgments is twofold. First, the assessment of adverse effects and drug safety, especially based on secondary analyses of RCT data, is a problem of increased public health concern and therefore makes this case study of timely interest. Second, by collecting and making the information from the relevant RCTs and their results publicly available, the FDA has helped facilitate the continued scientific and methodological developments that will help contribute to the monitoring of drug safety and the assessment of adverse effects.

2. Methods

2.1 General Approach

First we describe an approach for assessing the representativeness of RCT data for a target population. This methodology can be used when either individual level data is available from a trial (or trials) or when aggregate data are available as in a meta-analysis. Because every situation will be different we first identify general steps and the types of decisions that will typically need to be made in the implementation of this approach for making generalizability judgments. We then illustrate these steps in the context of the case study.

Step 1. Identifying Data Sources. For a given RCT or collection of RCTs, we must specify the question(s) of interest and identify the target population of interest. That is, to what population will the results of the RCTs be applied? It is then necessary to identify relevant data bases that are known to be representative of the population of interest or at least some identifiable sub-component of the population of interest. In our experience we have used nationally representative health surveys and epidemiologic studies. Large administrative data bases may prove useful too.

Step 2. Data Subsetting: Demographic Variables. Often there will not be a perfect match between the demographic variables in the trials with the demographic variables available from the outside data sources. For example, the RCTs may have included children and adolescents but the available nationally representative data bases may only have studied adolescents. In this case, it is necessary to use the subset of patients in the trials or the collection of trials that include only adolescents. By necessity this process will have restricted the target population of interest, for example, to adolescents instead of to children and adolescents. Another example is that there may be a possible mismatch between the specification of demographic variables in the RCTs versus the outside data sources, e.g., the specification of racial or income categories may not match up between the two. When this situation occurs, it will be necessary to group categories together or collapse categories to achieve as close a match as possible. These are examples of implementation decisions that arise in practice and need to be made explicit.

Step 3. Outcome Variable. In some cases the determination of the outcome of interest for both the RCTs and for the outside data sources may be straightforward, e.g., if the outcome of interest is mortality. In other cases, however, it may be necessary to construct an outcome measure that is as similar as possible among the different data sources. The case study below will illustrate such a situation. Just like in any data analysis many decisions will be made and it will be important to assess the sensitivity of the conclusions to those decisions.

Step 4. Sensitivity Analysis. When the basis for a data analytic decision is tenuous, it is important to check whether reasonable alternatives to that decision appreciably affect

the conclusions. In other words, it is important to check how sensitive the conclusions are to the method of analysis or to changes in the data. Sensitivity analysis is a systematic approach to address the question, "What happens if some aspect of the data or the analysis is changed?"

2.2 Case Study

Step 1. Identify Data Sources

We use two separate data sources in our analysis: the collection of randomized controlled treatment trials used by the FDA in their meta-analysis [13, 16], and the Youth Risk Behavior Survey conducted by the Centers for Disease Control and Prevention that is a representative probability sample of non-institutionalized adolescents in the U.S. [17].

Randomized Controlled Trials

At the request of the FDA, manufacturers identified all placebo-controlled trials of antidepressants conducted in children and adolescents, regardless of the indication studied, and provided information from these trials to the FDA. As noted earlier, this database consists of 24 randomized placebo-controlled efficacy trials of antidepressants in children and adolescents [13, 18]. Among the 24 pediatric placebo-controlled trials, 16 studied efficacy for Major Depressive Disorder (MDD) and 8 studied efficacy for other disorders. In this paper, we focus only on the adolescent MDD trials. Table 1 presents some characteristics of these studies (see also [19]).

Youth Risk Behavior Survey

For our target population, we sought a nationally-representative probability sample of depressed youth in the United States. The Youth Risk Behavior Survey (YRBS) is conducted every other year and is administered to students in grades nine through twelve [17]. We focus on the federally sponsored portion of the survey.

The national YRBS sample was designed so the survey results are representative of the behaviors of adolescents attending school throughout the United States. This was achieved by first forming counties and other municipalities into primary sampling units (PSUs). These PSUs were then stratified by urban character and size of black and Hispanic student populations. Then, PSUs were randomly selected from each strata, and schools were randomly selected from each chosen PSU. Finally, mandatory classes were selected from each nineth through twelfth grade level of each chosen school. 77% of the schools responded to the survey and 86% of the students responded to the survey, resulting in an overall response rate of 66%. Weighting methods were used to adjust the YRBS data to reduce the effects of the nonresponse. In our analysis we use the 1999 YRBS because it is the median year of recruitment commencement for the RCTs.

Drug Class	Drug Formulation	Number of Patients	Age Range	Exclude At Risk
SSRI	citalopram (Celexa)	178	7–17	yes
SSRI	citalopram (Celexa)	244	13–18	no
SSRI	paroxetine (Paxil)	206	7–17	yes
SSRI	paroxetine (Paxil)	275	13–18	yes
SSRI	paroxetine (Paxil)	181	12–18	yes
SSRI	fluoxetine (Prozac)	221	12–17	yes
SSRI	fluoxetine (Prozac)	96	8–18	no
SSRI	fluoxetine (Prozac)	219	8–17	yes
SSRI	fluoxetine (Prozac)	40	12–17	yes
SSRI	sertraline (Zoloft)	188	6–17	yes
SSRI	sertraline (Zoloft)	185	6–17	yes
Atypical	venlafaxine (Effexor)	165	7–17	yes
Atypical	venlafaxine (Effexor)	196	7–17	yes
Atypical	mirtazapine (Remeron)	259	7–17	yes
Atypical	nefazodone (Serzone)	278	7–17	yes
Atypical	nefazodone (Serzone)	190	12–17	yes

Table 1: Properties of Studies of MDD. Shaded rows indicate the studies compared with YRBS estimates. SSRI denotes Selective Serotonin Reuptake Inhibitor.

Step 2. Data Subsetting: Demographic Variables

We begin by taking subsets of the available data to accommodate the designs of both the RCTs and the survey. Because the YRBS only surveys adolescents, our analysis is limited to those RCTs that only studied adolescents aged 12 or older (see Table 1). Since the YRBS does not ask about psychological disorders such as obsessive-compulsive disorder, anxiety disorders or ADHD, our analysis is limited to the subset of adolescents from the YRBS with self-report of depressive symptoms (details below). Finally, because individual level data are not available from the trials, we do not know the demographic characteristics, even in the aggregate, of the adolescents in the trials that included both children and adolescents. Therefore, our primary comparative analysis is based on the 6 adolescent only MDD RCTs.

The first part of our analysis compares the demographic profiles of the youths included in the RCTs with the target population of United States depressed adolescents, as estimated from the YRBS. Taking the complex design of the survey into account using sampling weights, we create a demographic profile of American adolescents with depression based on age, gender and race (see, e.g., [20]). Since the YRBS is a survey of ninth through twelfth graders, we approximated as 12 years old all students who indicated they are "12 years old or younger" and approximated as 18 years old all students who indicated they are "18 years old or older". All the clinical trial investigators provided the racial demographics of their subjects according to four categories: white, black, Hispanic, and other. We re-coded the eight racial categories reported by the YRBS respondents to match these four categories such that the YRBS categories: "American Indian/Alaska Native", "Asian", "Native Hawaiian/Other PI", and "Multiple - Non-Hispanic", were defined as "Other" to correspond to the RCT categorization. To assess the demographic generalizability of the RCT results, we compare this estimated target population demographic profile with the estimated demographic profile of the population represented by the RCT studies. We make the comparison by calculating a confidence interval for the difference between the summary variables (means and proportions) in the two populations under standard normality assumptions.

The estimates of the demographic characteristics of the RCT population are constructed by combining the RCT samples and considering these subjects to be a random sample from an infinite study population. To estimate the variances of the estimates of the study population proportions, we use standard binomial variance formulas. Unfortunately, since individual data are not available for the trials, variance estimates were not available at the RCT study level for the age variables, we therefore conservatively use the maximum possible variance estimated by assuming all the subjects' ages consist of only the extreme values of 12 and 18 years old.

Step 3. Primary Outcome Variable

The second part of the analysis compares the suicidality rates in the two groups. For the RCTs, we use the FDA's definition of a suicidality event that was based on the

assessment and classification of adverse events conducted by an independent adjudication committee [13, 21]. According to the committee's scheme, an event is classified as suicidal if it is a suicide attempt, preparatory actions toward imminent suicidal behavior, or suicidal ideation. We note that there were no completed suicides in any of the RCT studies.

We identified four self-report questions from the YRBS that we will use (i) to identify a subset of youth in the population with depressive symptomatology and (ii) to identify those youth with suicidal ideations and/or behaviors. The question used to identify the target population of depressed youth is based on the SCID's stem question for depression [22]:

> Depression Screen. *During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?*

All students who responded positively to this screen question represent our target population and will be referred to as the "YRBS-Depressed" group. We note that Kessler, et al. [23] suggest that the rate of depression among adolescents who choose not to attend school is higher than those who do. With this in mind, it seems reasonable to assert that those who choose to drop out of school do not on average have less severe depression than those who stay in school.

We used the following YRBS questions to identify suicidal ideation and behavior in the sample of youths:

A. *During the past 12 months, did you ever seriously consider attempting suicide?*

B. *During the past 12 months, how many times did you actually attempt suicide?*

C. *If you attempted suicide during the past 12 months, did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?*

Question A from the YRBS is described by the National Center for Health Statistics as measuring suicidal ideation [17]; Questions B and C both indicate suicidal behavior. If a respondent endorsed at least one of items A-C, we consider that respondent to have had a suicidality event in the previous 12 months. We defer to the Appendix our analysis of the comparability of these self-report questions to the clinical assessments conducted in the RCTs.

When comparing the suicidality events between the two data sources, we balance the YRBS and RCTs according to gender, race and age to eliminate the possibility that these variables are actually the source of any differences we may find in the suicidality rates. We accomplish this by raking the YRBS data according to gender and race so that the YRBS proportions match the calculated RCT proportions. Raking is used in survey sampling for adjustments when a survey sample may cover segments of the target population in proportions that do not match the proportions of those segments in the population itself

(see, e.g., [20, pp. 269–271]). We used these post-stratified (raked) data to estimate the 12-month suicidality event rates for depressed American adolescents.

To compare the RCT and YRBS suicidality events, we also must match the time frame of the YRBS to the time frame of the RCTs. The RCTs lasted an average of 8 weeks, while the YRBS measured suicidality over a 12 month period. We approximated an eight week time period in the YRBS by dividing the 12 month rates by 6.5, assuming a uniform distribution of events over the 12 months. If this was not the actual distribution of events in the YRBS, there would be peaks and valleys in the distribution of events. Thus, the estimate of event rate would underestimate the peak time periods and overestimate the valley time periods. Because the patients enrolled in the RCTs are more likely to be in a peak than a valley time period [9], the 8-week YRBS suicidality rate, should if anything, underestimate the 8-week rate observed in the RCTs. We compare the rates estimated from the YRBS to the RCT rates using the same criteria as for the demographic comparisons.

Step 4. Sensitivity Analysis

We also created sub-groups of the RCT trials based on their exclusion of subjects with an elevated risk of suicide. First, we compared all the MDD RCT studies of adolescents to the "YRBS-Depressed" sample. Then, we compared the subset of MDD studies that excluded youths at high-baseline risk of suicidality, and finally we compared the one MDD study that did not have an exlusion for baseline suicidality to the age-adjusted "YRBS-Depressed" sample. In each case, we repeat the post-stratification process and compare each sub-group to the event rates of depressed adolescents, as estimated from the appropriately raked YRBS sample.

3. Results

Table 2 presents a comparison of the demographic characteristics for all the RCTs, the exclusion-based subsets of RCTs, and the "YRBS-Depressed" sample. Entries in bold face represent significant differences between the RCT and YRBS estimates. We note that the RCT subjects, on average, tend to be younger than the YRBS population (although not significantly so), and a significantly greater proportion of these subjects are white. The one RCT that did not exclude subjects at elevated risk of suicide had a higher proportion of female subjects than the YRBS estimate of the population proportion of depressed female adolescents in the U.S.

Figure 1 displays the adjusted 8-week suicidality rates for the RCTs studies and for the YRBS. Reading from left to right, the combined suicidality rate based on all 6 RCTs is 3.6% compared to 7.1%, the estimated rate among depressed U.S. adolescents, and is a highly statistically significant difference ($p < 0.001$; 95% CI for the difference in suicidality rates: [2.4, 4.6]). When we remove the one RCT that did *not* exclude adolescents at high risk of suicide and compare the five MDD RCTs that did exclude these children we have a similar statistically significant difference in suicidality rates, 3% versus 7.1% (95% CI

	Depressed American adolescents, as estimated from 1999 YRBS, ages 12-18	All MDD RCTs of adolescents n=6, N=1,151	MDD RCTs that excluded high risk n=5, N=907	MDD RCT that did not exclude high risk n=1, N=244
	estimate (standard error)	estimate (standard error)	estimate (standard error)	estimate (standard error)
Average Age	16.14 (0.04)	14.76 (2.99)	14.46 (2.95)	15.90 (2.86)
% Female	62.6 (1.8)	63.8 (1.4)	61.0 (1.6)	**74.0 (2.8)**
% White	54.1 (3.5)	**80.1 (1.2)**	**75.5 (1.4)**	**97.0 (1.1)**

Table 2: Demographic profiles of and tests of differences between depressed American adolescents aged 12 to 18 years old, and RCTs of MDD in adolescents aged 12 to 18 years old. n denotes the number of RCTs and N denotes the total number of subjects. Bold face entries represent significant differences between the RCT and YRBS estimates
*Variances calculated using the continuity correction (adding 1/2 to each race count in that study).

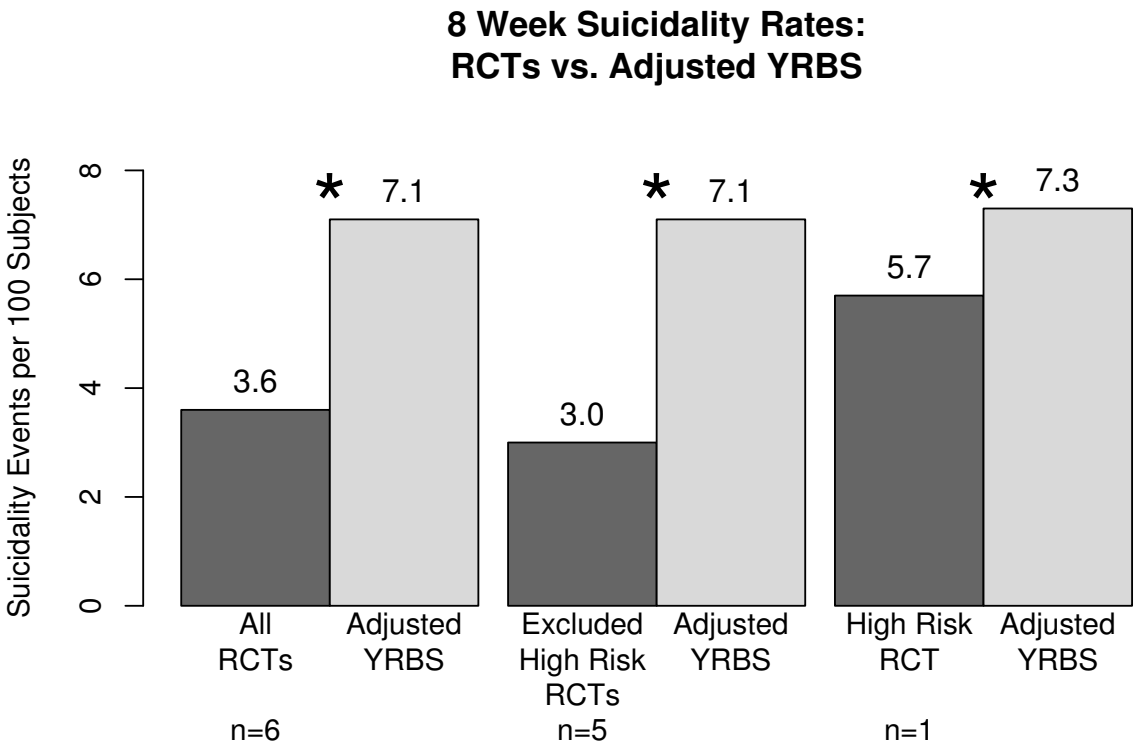

Figure 1: Comparison of the 8-week suicidality rate in the RCTs studies versus the age adjusted YRBS rate. Reading from left to right, the first comparison is of the 6 adolescent MDD RCTs; next is the subset of RCTs that exluded patients at high basline risk of suicidality; and finally, the one RCT that did not exclude high risk patients.

for the difference in suicidality rates: [2.9, 5.3]). Finally, when we compare the suicidality rate in the one RCT that included the high risk youth to the adjusted YRBS sample we still find a statistically significant difference but the magnitude of the difference as might be expected is smaller, 5.7% versus 7.3% (95% CI for the difference in suicidality rates: [0.9, 2.3]).

Because individual age data are not publicly available for the RCTs, as a sensitivity analysis for the effects of age differences between the two studies, we removed all the YRBS respondents aged 16-18 from the analysis to obtain a mean age for the YRBS data closer to that observed in the RCTs. The results (not displayed here) were nearly identifical to the above analysis which included the 16-18 year olds.

4. Discussion

4.1 Case Study

We compared the characteristics of the adolescents who participated in the MDD trials used in the FDA meta-analysis to characteristics of the general population of youth with depression. Our goal was to assess the representativeness of the RCT participants, and in particular, to assess whether there was a selection effect due to having excluded patients at high risk of suicidality from the trials. We found that the RCT subjects and the YRBS population were similar in terms of most demographic characteristics though a significantly greater proportion of the trial patients were white. Our major finding is that the rate of suicidal ideation and suicidal behaviors in the depressed adolescents who participated in the RCTs is approximately one-half the adjusted rate among depressed adolescents in the US as estimated by the YRBS.

What are the implications of this difference in suicidality rates? Since ideal, representative trials are not readily available we can only speculate on what these possible implications are. Consider a very large randomized trial of antidepressant treatment versus placebo-control in a truly representative sample from the target population of depressed adolescents where the outcome is suicidality. Let RR_{TP} denote the relative risk for suicidal events for treated versus control patients in the target population from this hypothetical study. Note that we would expect the base suicidality rate in such a trial to be greater than the observed rate estimated from the control groups in the actual trials used in the FDA meta-analysis. We consider three cases:

- *Case 1.* In the target population antidepressants *do* increase the risk of suicidality: $RR_{TP} > 1$.
- *Case 2.* In the target population antidepressants *do not* effect suicidality either negatively or positively: $RR_{TP} = 1$.
- *Case 3.* In the target population antidepressants *prevent* suicidal events: $RR_{TP} < 1$.

Case 1 is an example where the results from the ideal trial are consistent with the FDA's findings. That is, even though the trial data used in the FDA meta-analysis came almost exclusively from depressed youths who were at low baseline risk of suicidality, the results of the meta-analysis would also apply to the target population. If Case 1 were true, it would then follow that with an increase use of antidepressants in the general population, we would expect to see an increase in the number of suicidal events and ideations. Recent reports, however, have suggested that the opposite is the case at least for suicidal events (see, e.g., [10, 11, 7, 6]). Cases 2 and 3, in comparison to the FDA result, are examples of a negative interaction between treatment and the baseline risk of suicidality. If either of these cases were true, then the FDA estimate of risk would be a biased over estimate of the risk for the target population, due, we believe, to the unrepresentativeness of the patients in the trials included in the FDA meta-analysis. There does seem to be evidence for Case 3 based on observational studies [8, 9] and more recently [24, 25]. Since the current state-of-knowledge is insufficient to know with confidence which one of these cases is correct, clearly more research is required [26, 27]. Nevertheless, in the context of selection effects and heterogeneity of treatment response, this discussion suggests a possible explanation for the differences in results reported in the FDA meta-analysis and the results reported from the large observational studies. (See also [28]). (See the Appendix for further discussion of the comparability of the YRBS assessments of depression and suicidality to the assessments used in the RCTs.)

It is important to note that the FDA's analysis based on their meta-analysis of the RCTs did identify an approximate two-fold increased risk of suicidality among the children and adolescents who participated in those trials. From a regulatory perspective it is understandable why the FDA recommended a black box warning to improve treatment safety. Specifically, the FDA did not recommend against the use of antidepressants in pediatric treatment; its aim, rather, was to improve medical care by stimulating more extensive and close monitoring of children using antidepressants. These results, however, do suggest the need for further investigation into the clinical significance of suicidal ideation [29] and the need to identify risk factors for suicidality among pediatric antidepressant users [30].

4.2 General

RCTs are the gold standard for generating evidence about the efficacy of interventions, but to the extent that participating patients are not representative of the target population, their generalizability is limited. Similarly, the use of meta-analysis of RCTs is crucial for the accumulation and evaluation of evidence, especially concerning questions of safety, but combining information from a number of similar randomized trials often suffers from the *same* selection biases as evidence generated from a single randomized trial. This case study underscores the need to consistently assess the representativeness of study samples from randomized trials, especially in considering the evidence for treatment guidelines or drug approval for use in target community populations. The use of national

health survey data and/or epidemiology databases to evaluate the representativeness of randomized trial data is an important step in the translation of findings from academic and pharmaceutical company research to practice in the community.

As the practice of assessing generalizability becomes more common, it will become necessary to reach consensus on how "representative" samples should be and what criteria are necessary to determine whether a particular sample meets this goal. Moreover, such comparisons should stimulate clinical investigators to broaden their study inclusion criteria and limit their exclusion criteria to increase the generalizability of their findings, a recommendation recently made by the Institute of Medicine concerning the practice of excluding patients at risk of suicide from participation in clinical trials [31]. Until these changes are made, clinicians seeing patients in community settings will not have unequivocal evidence upon which to base treatment and safety decisions for life-threatening illnesses.

Acknowledgments

This study was funded in part by grants MH7862 from the NIMH and by The Pittsburgh Mind-Body Center, NIH grants HL076852/076858 (JBG).

Appendix

The validity of the argument comparing the suicidality rates in the RCTs to the target population is based on the comparability of the YRBS assessments of depression and suicidality to the assessments used in the RCTs. First, consider the YRBS question for depression. As noted earlier, this is the SCID stem question that in a clinical interview would lead to further diagnostic questions. Because the YRBS does not use the follow-up questions, the respondents who endorsed the YRBS question would include not only adolescents who are clinically depressed but also some who are not. Thus the group identified as YRBS-Depressed is a larger fraction of the population than those who would be identified for participation in the clinical trials. How would our results have changed if we had been able to restrict the YRBS sample to clinically depressed youths, similar to the RCT sample? In a hypothetical clinically-depressed YRBS subset, the rate would most likely have been higher than was found in the YRBS-depressed group in Figure 1, because severity of depression is associated with suicidality. Therefore, we believe that our estimate of the rate of suicidality in the YRBS-Depressed group is a lower bound to the suicidality rate for clinically depressed youths in the general population.

Next, consider the comparability of the suicidality assessments in the RCT's and the YRBS. It is possible that the methods of assessments in the RCTs might have detected more suicidality than the method of the YRBS: the assessment of suicidality in the YRBS was limited to three survey questions; the RCTs, however, often used both research instruments and adverse event logs. By using multiple methods, the suicidality assessments in the RCTs might have been more sensitive than the YRBS assessments. If this were the case, then the true difference between the suicidality rates in the population of depressed adolescents and the subset of that population that was enrolled in the RCTs would be even greater than we have reported here.

On the other hand, the method for detecting suicidality in the YRBS might have been more likely to detect suicidality than the methodology of the RCTs. Perhaps the anonymity of the YRBS, for example, made it more likely that respondents would endorse the suicidality items. In this case, the true differences between the suicidality rates in the depressed population and in the subset enrolled in the clinical trials would be smaller than reported here. Suppose for the moment that this is true. How much could this effect our conclusions? Recall that there was one RCT that did not exclude adolescents at high risk for suicide. The suicidality rate in this RCT was lower than that found among YRBS-depressed youths, the difference being 1.6% (see Figure 1). This difference may be due to many factors, including other selection biases. Let's assume, however, that the difference between the YRBS-depressed and the "no-exclusion" RCT rates is attributable entirely to the difference in the assessment techniques used in the YRBS and this RCT. We note that this difference, 1.6%, is not contained in the 95% confidence interval for the difference between the rate estimated from *all* of the trials versus the adjusted YRBS ([2.4, 4.6]). Thus, even under a worst-case assumption about the sensitivity associated with the varying suicidality assessment methods, the magnitude of the observed difference is still greater than can be explained by the (hypothetical) measurement artifact. In summary,

we do not know whether the assessments in the YRBS or the RCT were more likely to detect suicidality or whether they are both respectively unbiased. Whichever is the case, the data support a generalizability bias in the RCT results.

References

[1] Rothwell PM. Treating Individuals 1: External Validity of randomised controlled trials: "To whom do the results of this trial apply?". Lancet. 2005 January 1;365:82–93.

[2] Tunis SR, Stryer DB, Clancy CM. Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. JAMA. 2003 September 24;290(12):1624–1632.

[3] Altman D, Schultz K, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. *Annals of Internal Medicine*. 2001;134:663 – 694.

[4] Delate T, Gelenberg AJ, Simmons VA, Motheral BR. Trends in the use of antidepressants in a national sample of commercially insured pediatric patients, 1998 to 2002. Psychiatr Serv. 2004 Apr;55(4):387–391.

[5] Libby A, Brent D, Morrato E, Orton H, Allen R, Valuck R. Decline in Treatment of Pediatric Depression after FDA Advisory on Risk of Suicidality with SSRIs. *American Journal of Psychiatry*. 2007;164:884 – 891.

[6] Nemeroff C, Kalalil A, Keller M, Charney D, Lenderts S, Cascade E, et al. Impact of Publicity Concerning Pediatric Sucidality Data on Physician Practice Patterns in the United States. *Archives of General Psychiatry*. 2007;64:466 – 472.

[7] Hamilton B, Miniño A, Martin J, Kochanek K, Strobino D, Guyer B. Annual Summary of Vital Statistics: 2005. *Pediatrics*. 2007;119:345 – 360.

[8] Valuck RJ, Libby AM, Sills MR, Giese AA, Allen RR. Antidepressant treatment and risk of suicide attempt by adolescents with major depressive disorder: a propensity-adjusted retrospective cohort study. CNS Drugs. 2004;18(15):1119–1132.

[9] Simon GE, Savarino J, Operskalski B, Wang PS. Suicide risk during antidepressant treatment. Am J Psychiatry. 2006 Jan;163(1):41–47. Available from: `http://dx.doi.org/10.1176/appi.ajp.163.1.41`.

[10] Gibbons RD, Hur K, Bhawnik DK, Mann JJ. The relationship between antidepressant medication use and rate of suicide. *Archives of General Psychiatry*. 2005;62:165–172.

[11] Gibbons RD, Hur K, Bhaunik DK, Mann JJ. The relationship between antidepressant prescription rates and rate of early adolescent suicide. *American Journal of Psychiatry*. 2006;163:1898–1904.

[12] Kaizar EE, Greenhouse JB, Seltman H, Kelleher K. Do antidepressants cause suicidality in children? A Bayesian Meta-Analysis. Clinical Trials. 2006;3(2):73–90.

[13] Hammad TA. Review and Evaluaton of Clinical Data; 2004. Last accessed June 10, 2005. http://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4065b1-10-TAB08-Ha%mmads-Review.pdf.

[14] Longford NT. Selection bias and treatment heterogeneity in clinical trials. Stat Med. 1999 Jun;18(12):1467–1474.

[15] Stevens J, Kelleher K, Greenhouse J, Chen G, Xiang H, Kaizar E, et al. Empirical Evaluation of the Generalizability of the Sample from the Multimodal Treatment Study for ADHD. Adm Policy Ment Health. 2006 Oct;Available from: http://dx.doi.org/10.1007/s10488-006-0097-4.

[16] Hammad TA, Laughren T, Racoosin J. Suicidality in Pediatric Patients Treated With Antidepressant Drugs. *Archives of General Psychiatry*. 2006;63:332–339.

[17] Centers for Disease Control and Prevention. Methodology of the Youth Risk Behavior Surveillance System. MMWR. 2004;53(RR-12).

[18] Hammad TA. Results of the Analysis of Suicidality in Pediatric Trials of Newer Antidepressants; 2004. Last accessed June 10, 2005. http://www.fda.gov/ohrms/dockets/ac/04/slides/2004-4065S1_08_FDA-Hammad%.ppt.

[19] U S Food and Drug Administration. Characteristics of Pediatric Antidepressant Trials; 2004. Handout at the September advisory committees meeting. Last accessed June 10, 2005. http://www.fda.gov/ohrms/dockets/ac/04/briefing/2004-4065B1_29_Handout-Tables.pdf.

[20] Lohr SL. Sampling: Design and Analysis. Duxbury Press; 1999.

[21] Leslie LK, Newman TB, Chesney PJ, Perrin JM. The Food and Drug Administration's Deliberations on Antidepressant Use in Pediatric Patients. Pediatrics. 2005;116:195 – 204.

[22] Spitzer R, Williams J, Gibbon M, First M. *Structured Clinical Interview for DSM-III-R, Patient Edition/Non-patient Edition,(SCID-P/SCID-NP)*. Washington, D.C.: American Psychiatric Press, Inc.; 1990.

[23] Kessler RC, Foster CL, Saunders WB, Stang PE. Social Consequences of Psychiatric Disorders I: Educational Attainment. American Journal of Psychiatry. 1995 July;152(7):1026–1032.

[24] Simon G, Savarino J. Suicide Attempts among Patients Starting Depression Treatment with Medications or Psychotherapy. *American Journal of Psychiatry*. 2007;164:1029 – 1034.

[25] Gibbons R, Brown H, Hur K, Marcus S, Bhaunik D, Mann J. Telationship between antidepressants and Suicide attempts: An Analysis of the Veterans Health Administrative Data Sets. *American Journal of Psychiatry*. 2007;164:1044–1049.

[26] Greenhouse JB, Kelleher KJ. Thinking outside the (black) box: Antidepressants, suicidality, and research synthesis. *Pediatrics*. 2005;116:231 – 233.

[27] Simon G. How can we know whether antidepressants increase suicide risk? *American Journal of Psychiatry*. 2006;163:1861 – 1862.

[28] Zimmerman M, Chelminski I, Posternak M. Generalizability of Antidepressant Efficacy Trials: Differences between Depressed Psychiatric Outpatients Who Would or Would Not Qualify for an Efficacy Trial. *American Journal of Psychiatry*. 2005;162:1370 – 1372.

[29] Klein D. The Flawed Basis for FDA Post-Marketing Safety Decisions: The Example of Anti-Depressants and Children. *Neuropsychopharmacology*. 2006;31:689 – 699.

[30] Bridge J, Iyengar S, Salary C, Barbe R, Birmaher B, Pincus H, et al. Clinical Response and Risk for Reproted Suicidal Ideation and Suicide Attempts in Pediatric Antidepressnat Treatment: A Meta-analysis of Randomized Clinical Trials. *Journal of the American Medical Association*. 2007;297:1683 – 1696.

[31] Committee on Pathophysiology and Prevention of Adolescent and Adult Suicide, Board on Neuroscience and Behavioral Health. *Reducing Suicide: A National Imperative*. Goldsmith S, Pellmar T, Kleinman A, Bunney W, editors. National Academies Press; 2002.

Figure Legends

Figure 1: Comparison of the 8-week suicidality rate in the RCTs studies versus the age adjusted YRBS rate. Reading from left to right, the first comparison is of the 6 adolescent MDD RCTs; next is the subset of RCTs that exluded patients at high basline risk of suicidality; and finally, the one RCT that did not exclude high risk patients.