Generalizability of Studies on Mental Health Treatment and Outcomes, 1981 to 1996

Joel T. Braslow, M.D., Ph.D.
Naihua Duan, Ph.D.
Sarah L. Starks, B.A.
Antonio Polo, Ph.D.
Elizabeth Bromley, M.D., M.A.
Kenneth B. Wells, M.D., M.P.H.

Objective: **This study operationalized and measured the external validity, or generalizability, of studies on mental health treatment and outcomes published in four journals between 1981 and 1996. *Method:* MEDLINE was searched for articles on mental health treatment and outcomes that were published in four leading psychiatry and psychology journals between 1981 and 1996. A 156-item instrument was used to assess generalizability of study findings. *Results:* Of more than 9,000 citations, 414 eligible studies were identified. Inclusion of community sites and patients from racial or ethnic minority groups were documented in only 12 and 25 percent of studies, respectively. Random or systematic sampling methods were rare (3 percent), and 75 percent of studies did not explicitly address sample representativeness. Studies with funding from the National Institute of Mental Health (NIMH) were more likely than those without NIMH funding to document the inclusion of patients from minority groups (30 percent compared with 20 percent). Randomized studies were more likely than nonrandomized studies to document the inclusion of patients from minority groups (28 percent compared with 17 percent), include patients with comorbid psychiatric conditions (31 percent compared with 19 percent), and attend to sample representativeness (28 percent compared with 15 percent). Modest improvements were seen over time in inclusion of patients from minority groups, inclusion of patients with psychiatric comorbidities, and attention to sample representativeness. *Conclusions:* Generalizability of studies on treatments and outcomes, whether experimental or observational, remained low and poorly documented over the 16-year period. (*Psychiatric Services* 56:1261–1268, 2005)**

Dr. Braslow, Dr. Duan, and Dr. Wells are affiliated with the department of psychiatry and biobehavioral sciences at University of California, Los Angeles. Dr. Braslow is also affiliated with and the Mental Illness Research, Education, and Clinical Center at the Department of Veterans Affairs, VISN 22, in Los Angeles. Dr. Duan is also affiliated the department of biostatistics at the University of California, Los Angeles. Dr. Wells is also affiliated with RAND in Santa Monica. Ms. Starks is a doctoral student in the department of health services, Dr. Polo is with the department of clinical psychology, and Dr. Bromley is with the Robert Wood Johnson Clinical Scholars Program at the University of California, Los Angeles. Send correspondence to Dr. Braslow at the University of California, Los Angeles, 10920 Wilshire Boulevard, Suite 300, Los Angeles, California 90024 (e-mail, jbraslow@ucla.edu).

Despite substantial investment in research on clinical outcomes, little is known about the generalizability of this research to usual-care settings or its appropriateness as a guide to clinical practice (1). Concerns about the generalizability, or external validity, of the clinical research base have increased in response to the growing movement toward evidence-based medicine (2–6). This article addresses these concerns, especially the lack of systematic evaluations of the generalizability of studies on mental health outcomes, by examining outcomes research reported in leading psychiatric and psychological journals from 1981 to 1996.

This time frame was chosen to establish a baseline measure of external validity, which is usually defined as how well study findings apply to, or can be replicated with, other patient populations, providers, and treatment settings (7–12). We expect that follow-up research will show improvements over this baseline as a result of changes in research policy that occurred in the mid-1990s. The NIH (National Institutes of Health) Revitalization Act of 1993 required that all NIH-funded human subjects research (biomedical or behavioral) include women and persons from racial or ethnic minority groups. The act also required that phase III clinical trials achieve sufficient inclusion to allow for valid statistical analysis of such group differences in intervention effect. A few years later, the National Advisory Mental Health Council of the National Institute of Mental Health (NIMH) established the clinical treatment and services research workgroup; the yearlong efforts of the workgroup resulted in 49 specific recommendations that were published in the landmark 1998 *Bridging Science and Service* report (13).

A number of these recommendations focused on expanding the NIMH portfolio "in the domains of efficacy, effectiveness, practice, and service systems research." Of particular interest to health services researchers, the workgroup saw a need for "research that assesses the generalizability of interventions across diagnostic complexities (for example, comorbidity or chronicity), as well as individual, social, and demographic factors." Underlying all these recommendations is the recognition that clinical science needs to shed light not only on an intervention's efficacy but also on how well efficacious interventions actually work in diverse clinical settings, provider and patient populations, and practice circumstances.

Our review sought to assess how well the mental health literature addressed external validity in the years leading up to the *Bridging Science and Service* report, to identify particular areas in need of change, and to provide a baseline against which to assess future changes in the generalizability of clinical science. We examined how studies of psychiatric treatment and outcomes in four leading psychiatry and psychology journals tackled the issue of external validity over a 16-year period, 1981 through 1996. We evaluated the data for change over time, predicting improvements in reported external validity as a result of increased focus on the translation of research findings into everyday clinical practice and the need to report aspects of clinical trials that relate to generalizability (14–17). We also explored the differences between NIMH-funded and non–NIMH-funded studies, expecting that NIMH funding might yield higher external validity because of rigorous peer review and clearer science policy guidelines. Finally, we assessed the differences between randomized and nonrandomized studies in order to test the often-made, though unproven, assertion that observational studies have greater external validity than randomized clinical trials (18,19), whereas randomized studies have greater internal validity.

Methods

Part I: sampling frame and procedure and eligibility criteria

We restricted ourselves to the two leading journals in psychiatry—*American Journal of Psychiatry (AJP)* and *Archives of General Psychiatry (AGP)*—and psychology—*Journal of Consulting and Clinical Psychology (JCCP)* and *Journal of Abnormal Psychology (JAP)*—because these journals reflected the highest scientific standards in their respective fields during the study period. We selected articles between 1981 and 1996 so that our sample would reflect the state of research in the years leading up to the creation of the National Advisory Mental Health Council's clinical treatment and services research workgroup. Although randomized clinical trials were introduced into clinical science in the 1950s, it was not until the 1960s and 1970s that they became standardized and concepts of external validity were well articulated. Furthermore, the distinction between efficacy and effectiveness—closely linked to internal and external validity, respectively—did not appear in print until 1971 (14).

> *One of the goals of our review was to assess how well the mental health literature addressed external validity in the years leading up to the Bridging Science and Service report.*

We included all articles that examined the effect of a treatment for a mental health condition in comparison with a control group. We defined treatment broadly to include any intervention given with the intent to improve outcome. Eligible treatments included psychotropic medications, electroconvulsive therapy, psychotherapies, psychosocial interventions, and practice-based quality improvement interventions and policies. Eligible outcomes included clinical outcomes and course, psychological symptoms, morbidity and functioning, satisfaction with care, and cost-related measures. The following methods were eligible: concurrent controls, whether derived experimentally (for example, patients randomized into two study arms) or observationally (for example, case management compared with usual care within a clinic setting); historical controls (for example, patient outcomes compared with historical outcomes from available datasets); and self as control (for example, patients first taking medications and then taking placebos). We excluded meta-analyses and literature reviews.

We excluded studies that focused on side effects rather than effectiveness or efficacy and studies that examined predictors or moderators of treatment response—for example, whether depression outcomes with antidepressant treatment are predicted by whether a patient has borderline personality traits. We also excluded studies that examined treatment of the following disorders: physical health conditions without a concurrent mental disorder, substance-related disorders without a concurrent mental disorder, mental disorders caused by a medical condition, sleep disorders, mental retardation, cognitive disorders (for example, dementia, delirium, and organic brain syndrome), and factitious disorders. However, we did include studies in which patients had comorbid condi-

Table 1

Definition of key variables used to assess the generalizability of studies on mental health treatment and outcomes

Variable	Definition
Article characteristics	
Journal title	*American Journal of Psychiatry*, *Archives of General Psychiatry*, *Journal of Consulting and Clinical Psychology*, and *Journal of Abnormal Psychology*
Year published	Continuous, 1981 through 1996
Study characteristics	
Study design	Parallel cohort design (interventions given to separate cohorts) compared with crossover design (given to same cohort at different times)
Randomized treatment allocation	Treatment randomly allocated (experimental study design)
Interventions	Medication or other somatic, psychosocial, or combined
Treatment settings	Outpatient, inpatient, or other or unknown
Funding sources	Funding sources include the National Institute of Mental Health
Context, providers, and implementation	
Community site included	Inclusion of a Department of Veterans Affairs hospital or a site that has no academic affiliation, in addition to or instead of an academic site
Usual-care or other community providers	Treatment provided either by usual-care providers or another community provider, in addition to or instead of the study team
Flexible treatment protocols	Flexibility permitted by study protocol in the use of interventions
Patient characteristics	
Persons from racial or ethnic minority groups included	Study reported inclusion of nonwhites
Women included	Study reported inclusion of women
Comorbid psychiatric conditions included	Study reported inclusion of patients with at least one comorbid psychiatric condition
Comorbid medical conditions included	Study reported inclusion of patients with at least one comorbid medical condition
Children included	Study reported inclusion of children (younger than 18 years)
Sampling methods	
More than one sampling level	Study reported sampling at a level higher than the individual patient (for example, clinic, hospital, and city)
Overall sampling strategy	Sampling strategy used to select the study population (or, if multiple levels sampled, the "best" strategy used by the study). In order of decreasing preference: random or universal, systematic, quota, and convenience
Sample representativeness attended to	Whether the study reported one or more of the following: the number of patients approached for the study, the refusal rate, consideration of refusal bias, or weighting of the sample to the parent population

tions (for example, a primary mental disorder with a co-occurring substance use disorder), and we included studies for all age groups.

On the basis of preliminary screening, we divided the four journals into two groups by the relative number of mental health treatment and outcome studies published. For the low-yield journal (*JAP*) we sampled the entire time period from 1981 to 1996, and for the high-yield journals (*AGP*, *AJP*, and *JCCP*) we instead sampled three periods: 1981 through 1983, 1988 through 1990, and 1994 through 1996. These periods were chosen to allow us to detect change over the entire period and to determine whether the trend was linear. We then obtained all records from MEDLINE for each journal over the selected periods and downloaded them into Procite for Windows, version 3.4. This approach yielded a total of 9,100 records. On the basis of the MEDLINE "publication type" field, we excluded citations in the following categories: letters, historical reports and biographies, comments, and editorials. This exclusion process left a total of 7,095 citations, 408 of which met our full inclusion criteria. Four of the articles contained two separate studies each, and one article contained three separate studies. In all, we had 414 eligible treatment and outcome studies.

Part II: abstraction instrument

Our goal was to operationalize and measure external validity, or generalizability, which usually is defined as how well study findings apply to, or can be replicated in, other patient populations, providers, and treatment settings (7–12). External validity depends on how well the study mirrors or represents community practice in terms of setting (for example, outpatient compared with inpatient), context (for example, academic compared with community), providers (for example, usual-care providers compared with researchers), treatment implementation (for example, rigid compared with flexible protocols), and patient characteristics (for example, age, gender, race or ethnicity, and comorbid medical and psychiatric conditions). Various sampling methods were used to ensure representativeness. Random sampling, on which nearly all inferential statistics depend, is considered statistically the most valid, followed (in descending order of preference) by systematic, quota, and convenience or volunteer

Table 2

Overall level and time-trend analysis of the generalizability of 414 studies on mental health treatment and outcomes

	Overall levels		Time-trend analysis (predicted values)					
	1981 to 1996		1981		1996		Linear time trend	
Variable	N	%	%	SD	%	SD	t†	p
Context, providers, and implementation								
Community site included	48	12	14	4	10	2	–.68	.5
Usual or other community providers	53	13	11	3	14	3	.54	.592
Flexible treatment protocols	190	46	42	5	49	4	.97	.333
Patient characteristics								
Persons from racial or ethnic minority groups included	105	25	15	3	34	4	2.9	.004
Women included	351	85	85	4	85	3	.11	.91
Comorbid psychiatric conditions included	116	28	19	4	35	4	2.35	.019
Comorbid medical conditions included	25	6	4	2	8	2	1.25	.213
Sampling methodology								
Random, universal, systematic, or quota	11	3	4	2	2	1	.9	.371
More than one sampling level	121	29	24	4	33	4	1.37	.172
Sample representativeness attended to	103	25	14	3	34	4	3.04	.003

†df=413

sampling. Studies with strong external validity employ representative sampling methods not only to individual patients but to providers and settings as well.

To obtain data on study characteristics and methods relevant to external validity, we developed an abstraction instrument with 156 individual items, from which we constructed the indices of external validity defined in Table 1. These indices assessed use of community sites, usual-care providers, and flexible protocols (for example, whether dosages could be adjusted in response to clinical changes). The indices also assessed the inclusion of persons from racial or ethnic minority groups, women, and patients with comorbid conditions. Also examined were the number of sampling levels reported (for example, city, clinic, provider, and patient) and the sampling methods employed (for example, random, systematic, quota, and convenience), both for the study as a whole and for each sampling level. Finally, the indices assessed the extent to which authors addressed the representativeness of the sample (for example, by assessing and mitigating nonresponse bias). Many of these indices were constructed from multiple variables. For example, in constructing the variable reported as "intervention," we coded 13 items for each study group, including whether the group was treated with each of the following: medication, electroconvulsive therapy, psychotherapy, rehabilitation, clinical case management, structured treatment assessment with minimal intervention or management, a financial intervention, quality improvement, or other. These variables were combined and reported as the values shown in Table 1.

Most of the variables in the instrument had three possible values: present, absent, or not reported. We reasoned that only reported positive data serves as a useful guide to community practitioners in weighing the external validity of a particular study. For this reason, we focused on explicit documentation of external validity, grouping absent and not reported into a single category. However, to illustrate the extent to which the lack of reporting is a problem in itself, we determined the percentage of studies that failed to report any data for key variables.

Seven psychology graduate students and one psychiatry resident used the instrument to abstract the 414 studies, which were distributed randomly by date and journal, both across reviewers and in abstraction sequence for each reviewer. Forty-eight studies (12 percent) were coded by two or more research assistants in order to assess mixed (interrater and intrarater) reliability by using variance component analysis (20,21). Nine variables had excellent interrater reliability with reliability coefficients ranging from .81 to 1.0: study design, use of medication or somatic interventions, use of psychosocial interventions, funding source, journal, inpatient setting, inclusion of persons from racial or ethnic minority groups, inclusion of women, and inclusion of children. Six variables had only moderate reliability, with coefficients ranging from .40 to .60: outpatient setting, inclusion of community sites, inclusion of usual-care or other nonresearch providers, flexible treatment protocols, comorbid psychiatric conditions, and comorbid medical conditions.

A final group of variables proved too difficult for the abstractors to code reliably: whether treatment allocation was randomized, the number of sampling levels, the sampling method, and attention to sample representativeness. We had the most experienced research assistant, the third-year psychiatry resident, recode these, including ones he had previously abstracted. One of the authors (JTB)

Table 3

Key variables of 414 studies on mental health treatment and outcomes in relation to funding by the National Institute of Mental Health (NIMH) and experiment type

	NIMH funding				Random allocation			
	Yes (N=216)		No (N=198)		Yes (N=314)		No (N=100)	
Variable	N	%	N	%	N	%	N	%
Context, providers, and implementation								
Community site included	21	10	27	14	35	11	13	13
Usual or other community providers	25	12	28	14	36	11	17	17
Flexible treatment protocols	95	44	95	48	142	45	48	48
Patient characteristics								
Persons from racial or ethnic minority groups included[a]	65	30	40	20	88	28	17	17
Women included	189	88	162	82	269	86	82	82
Comorbid psychiatric conditions included[b]	56	26	60	30	97	31	19	19
Comorbid medical conditions included	18	8	7	4	20	6	5	5
Sampling methods								
Random, universal, systematic, or quota	7	3	4	2	7	2	4	4
More than one sampling level	63	29	58	29	92	29	29	29
Sample representativeness attended to[c]	51	24	52	26	88	28	15	15

[a] t=2.52, df=413, p=.012 for NIMH funding; t=1.99, df=413, p=.048 for random allocation

[b] t=2.09, df=413, p=.037 for random allocation

[c] t=2.35, df=413, p=.019 for random allocation

checked all the recoded variables by comparing the research assistant's determination with the article itself and then discussing and resolving any uncertainties in consultation with the other authors (KBW and ND). Interrater reliability of those items was not reported, because a single individual reassessed them all and disagreements were resolved by consensus.

Part III: analysis

We employed linear time-trend analysis using a logistic regression model to test the effect of time (year) with the Wald t test. To demonstrate the magnitude of the change over time, we used the fitted logistic regression model to derive predicted values for 1981 and 1996. We expanded the model to examine differences between NIMH- and non–NIMH-funded studies and between studies in which treatment was randomly allocated and those in which it was self-selected (that is, experimental compared with observational or quasi-experimental studies), both controlled for the effect of time.

Results

The 414 studies in our sample were distributed among the four journals as follows: *AJP*, 152 studies, or 37 percent; *AGP*, 130 studies, or 31 percent; *JCCP*, 111 studies, or 27 percent; and *JAP*, 21 studies, or 5 percent. They covered a range of designs, interventions, and settings. Most studies used a parallel cohort design (324 studies, or 78 percent) rather than a crossover design (90 studies, or 22 percent), and most randomly allocated treatment (314 studies, or 76 percent) rather than allowing self-selection (100 studies, or 24 percent). More than half the studies tested medications or somatic interventions only (213 studies, or 51 percent), about one-third tested psychosocial interventions only (132 studies, or 32 percent), and 69 studies (17 percent) examined combined treatments. Within these three categories, the following treatments were examined: medications (268 studies, or 65 percent), electroconvulsive therapy (14 studies, or 3 percent), psychotherapy (163 studies, or 39 percent), rehabilitation (15 studies, or 4 percent), clinical case management (11 studies, or 3 percent), structured clinical management (five studies, or 1 percent), quality improvement (two studies, less than 1 percent), financial interventions (one study, less than 1 percent), and other psychosocial treatments (38 studies, or 9 percent). Outpatient settings were used in 288 studies (70 percent), and inpatient settings were used in 109 studies (26 percent). More than half (216 studies, or 52 percent) had NIMH funding. Sixty-nine studies (17 percent) included children.

As shown in Table 2, overall the 414 studies scored quite low on most of our measures of external validity. Most studies included women (85 percent), but only 25 percent reported any inclusion of nonwhites. Reported inclusion of comorbid conditions also was low: less than one-third of studies included patients with comorbid psychiatric conditions (28 percent), and an even smaller percentage enrolled patients with comorbid medical conditions (6 percent). Nearly half the studies (46 percent) used flexible treatment protocols, but only 12 percent were performed in community sites and only 13 percent relied on usual-care or nonresearch providers. Sampling methods also were inadequate: 3 percent reported employing random, universal, systematic, or quota sampling methods, whereas 167 (40

Table 4

Documentation of key variables used to assess the generalizability of 414 studies on mental health treatment and outcomes

	Did not report		Reported "no"[a]	
Variable	N	%	N	%
Treatment randomly allocated	37	9	63	15
Funded by the National Institute of Mental Health	95	23	103	25
Flexible treatment protocols	162	39	62	15
Persons from racial or ethnic minority groups included	296	71	13	3
Women included	30	7	33	8
Comorbid psychiatric conditions included	149	36	149	36
Comorbid medical conditions included	188	45	201	49
Children included	25	6	320	77
Sampling strategy other than convenience	236	57	167	40
More than one sampling level	100	24	196	47
Setting (outpatient, inpatient, or mixed)	236	57	na	na

[a] For example, reported that treatment was not randomly allocated or that persons from minority groups were not included

percent) reported using a convenience sample and 236 (57 percent) failed to report on the sampling method; only 25 percent reported efforts to address sample representativeness; and only 29 percent mentioned any sampling strategy for levels other than at the individual patient level (that is, provider, hospital, or region).

Change over time was significant in three areas (Table 2). First, predicted inclusion of persons from racial or ethnic minority groups increased from 15 percent of studies in 1981 to 34 percent in 1996. Second, predicted inclusion of patients with comorbid psychiatric conditions increased from 19 percent in 1981 to 35 percent in 1996. Finally, the predicted percentage of studies reporting attention to sample representativeness increased from 14 percent in 1981 to 34 percent of studies in 1996.

Differences based on NIMH funding and random allocations also were found. As shown in Table 3, across the study period NIMH-funded studies were significantly more likely than studies funded by other sources to include persons from racial or ethnic minority groups when the analyses controlled for both time and form of treatment allocation (30 percent compared with 20 percent). Also shown in Table 3, studies that randomly allocated treatments were significantly more likely than those that did not use random allocation to report that they had included persons from racial or ethnic minority groups (28 percent compared with 17 percent) and patients with comorbid psychiatric conditions (31 percent compared with 19 percent), as well as to have given some attention to sample representativeness (28 percent compared with 15 percent).

We emphasize that these results combine "not reported" with "reported no." In other words, we are scoring documented evidence for external validity. Table 4 shows the breakdown of studies without demonstrated external validity into the number and percentage for which the variables were "not reported" or "reported no." In a substantial proportion of studies, lack of documented external validity was due to nonreporting. For example, among the 414 studies, 71 percent did not report on race or ethnicity and 3 percent reported that persons from racial or ethnic minority groups were not included. Similarly, 57 percent did not report any sampling strategy, and 40 percent reported using a convenience sample.

Discussion

Many have suggested that the literature on mental health treatment and outcomes has paid scant attention to external validity. Our study of the literature provides strong empirical support for this belief and provides a baseline against which to measure efforts to improve external validity subsequent to the *Bridging Science and Services* report. As our study underscores, our field has fallen short in documenting, and most likely testing for, external validity, a necessary precondition for making valid generalizations. Twelve percent of the studies used community-based settings and 13 percent used usual-care providers, although most mental health treatments are received and delivered in community-based settings. A majority of studies did not report racial or ethnic minority distribution. Furthermore, most studies did not report sampling strategies used to select research sites, treatment settings, or patients. Random sampling, considered by most statisticians to be the best way to achieve a representative sample, was a rarity. More disconcerting, sampling bias received no attention in analysis or discussion in a majority of studies.

Also of concern was the difficulty that our coders had in reliably abstracting a number of key variables of the study design: whether treatment allocation was randomized (that is, experimental as opposed to quasi-experimental or observational), the number of sampling levels, whether a sampling method other than a convenience strategy was employed, and whether the study attended to sample representativeness. These are fundamental aspects of the methods used in studies of treatment and outcome, and as such they should be readily evident to readers of an article. Similarly, the moderately reliable variables (outpatient setting, inclusion of community sites, inclusion of usual-care or other nonresearch providers, flexible treatment protocols, and comorbid psychiatric and medical conditions) are important factors in allowing a clinician to assess whether a study applies to his or her patient population. The difficulty that we had in coding these variables, combined with significant nonreporting for some variables, highlights the challenge faced by readers wishing to understand the validity and generalizability of many of these studies.

However, there were a few bright spots. Mirroring treatment by nonresearch clinicians, treatment protocols with individual tailoring were used in a substantial percentage of studies. Furthermore, a large majority of studies included women as well as men, although we did not examine the proportion of women in these samples. Finally, the time-trend analysis suggests that the situation may be improving slightly when it comes to attention to sample representativeness and inclusion of persons from racial or ethnic minority groups and patients with comorbid psychiatric conditions. Nonetheless, there still is room for considerable improvement.

Contrary to our expectations, studies with random allocation had stronger generalizability in some domains than observational and quasi-experimental studies, at least for the inclusion of persons from racial or ethnic minority groups and patients with comorbid psychiatric conditions and for the attention to sample representativeness. Furthermore, we found few differences between studies with and without NIMH funding. Because NIMH policy requires the inclusion of persons from racial or ethnic minority groups, we expected that NIMH-funded studies would be more likely to report that they included persons from these groups, and indeed they were. However, inclusion of these groups was still low (30 percent of NIMH-funded studies compared with 20 percent of studies funded by another source). Most studies published by 1996, the end of our study period, most likely were generated and funded before the implementation of the NIH Revitalization Act of 1993. Thus inclusion of persons from racial or ethnic minority groups in NIMH-funded studies is likely to improve substantially over the baseline established by our results.

Our study has a number of limitations. First, we reviewed some of the most scientifically rigorous clinical journals; thus our conclusions may be conservative relative to the field as a whole. Second, our study covers only the period through 1996. A similar study should be conducted for the subsequent ten-to-15-year period to determine whether policies put in place to achieve greater generalizability accomplished this goal. Given the lead time to develop and publish studies, ten to 15 years seems to be a minimum period to identify such a shift in science policy. Third, a number of the variables in our abstraction instrument required a substantial amount of judgment (as illustrated by our difficulty in achieving satisfactory reliability on some key variables). Fourth, in our effort to assess the field broadly, we combined studies covering a range of illnesses, developmental stages, and treatment modalities. More or less progress might have occurred in specific specialty fields. Fifth, we focused on relatively crude indicators of generalizability—that is, presence or absence of a given domain as documented in a given article.

We urge investigators to adequately document and report on factors relevant to external validity and suggest that journal editors support the reporting of these factors.

Finally, although negative scores on our indices suggest a lack of documented external validity, they do not necessarily mean that a study's findings are actually biased for real-world applications. For example, if a particular treatment's effects are independent of the characteristics of patients, providers, treatment implementation, and settings, then a study that scores low on our external validity indices might still produce broadly applicable findings. However, assumptions of independence may be wrong, especially for mental health remedies in which the interventions themselves, as in the case of psychotherapy, often explicitly depend on the patient, provider, or context.

Conclusions

It appears that most of the studies published during this 16-year period were conducted with locations, circumstances, and patients that differ from the everyday clinical world to which their results are intended to be applied. Even if our findings were a reflection of poor reporting rather than actual omission, lack of documented external validity poses serious limitations for persons who wish to apply research findings to everyday clinical practice with some degree of confidence. Specifically, these deficiencies in external validity create uncertainty for the meaning and significance of data on clinical outcomes, potentially leading to inappropriate application of treatments, underuse of treatments that might have broad effectiveness, or arbitrary and uneven practices.

More fundamentally, the research portfolio we reviewed suggests that during this period internal validity might have been viewed as both a necessary and sufficient standard for guiding clinical practice. During this period clinical research in many fields had at best a crude approach to defining, measuring, evaluating, or enhancing external validity. It is our hope that the recommendations of the *Bridging Science and Service* report and the shift in priorities and resources that these recommendations represent, will help to advance the science of external validity as well as the extent to which external validity is actually achieved in studies.

We see four potential ways to improve external validity and its documentation. First, we urge investigators to adequately document and report on factors relevant to external validity (for example, sampling methods, patient and provider characteristics, and study site characteristics) and suggest that journal editors support the reporting of these factors, much as they have supported the reporting of internal validity variables. Even in trials in which external validity is less important (for example, phase II trials), it is important that readers be able to determine from an article the extent to which its results can be generalized.

Second, we suggest convening an expert panel to devise a checklist of items critical to the evaluation and documentation of external validity, to be used as a publication standard for treatment and outcome studies. Such a checklist could complement the Consolidated Standards of Reporting Clinical Trials (CONSORT) statement, a checklist that investigators can follow when reporting results of randomized clinical trials but that focuses primarily on internal validity and contains only a few external validity items (22,23).

Third, we recommend that researchers make a greater effort to study patients and contexts to which they wish to extend their findings. For example, investigators might employ random or systematic sampling rather than convenience sampling of the parent or target population and recruit usual-care providers and persons from racial or ethnic minority groups to better represent the population to which they hope to generalize.

Fourth, we recommend that researchers design studies that mirror how interventions will be used in actual clinical practice. These steps will require new standards, enriched methods, and training of clinical scientists to employ them. It is hoped that these steps will result in clinical trials that tell us more about how treatments will work in actual practice or how effects vary with context. ♦

Acknowledgments

Dr. Braslow was supported by a National Institute of Mental Health Career Development Award (K23-MH-01856) and by the Burroughs Wellcome Fund. The project also was supported by two National Institute of Mental Health research centers: UCLA-RAND Research Center on Managed Care for Psychiatric Disorders (grant P-50-MH-546-23) and the UCLA-RAND Research Center for Research on Quality in Managed Care (grant P-30-MH-0686-39). The authors thank Diana Liao, M.P.H., for data analysis and Hisham Korraa, M.D., for data collection.

References

1. Feinstein AR, Horwitz RI: Problems in the "evidence" of "evidence-based medicine." American Journal of Medicine 103:529–535, 1997
2. Bensing J: Bridging the gap: the separate worlds of evidence-based medicine and patient-centered medicine. Patient Education and Counseling 39:17–25, 2000
3. Carne X, Arnaiz JA: Methodological and political issues in clinical pharmacology research by the year 2000. European Journal of Clinical Pharmacology 55:781–785, 2000
4. Herman J: Shortcomings of the randomized controlled trial: a view from the boondocks. Journal of Evaluation in Clinical Practice 4:283–286, 1998
5. Tunis SR, Stryer DB, Clancy CM: Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 290:1624–1632, 2003
6. Upshur R, VanDenKerhof EG, Goel V: Meaning and measurement: an inclusive model of evidence in health care. Journal of Evaluation in Clinical Practice 7:91–96, 2001
7. Campbell DT: Factors relevant to the validity of experiments in social settings. Psychological Bulletin 54:297–312, 1957
8. Campbell DT, Stanley JC: Experimental and quasi-experimental designs for research on teaching, in Handbook of Research on Teaching: A Project of the American Educational Research Association. Edited by Gage NL. Chicago, Rand McNally, 1963
9. Jüni P, Altman DG, Egger, M: Systematic reviews in health care: assessing the quality of controlled clinical trials. British Medical Journal 323:42–46, 2001
10. Heinrichs RW: In Search of Madness: Schizophrenia and Neuroscience. New York, Oxford University Press, 2001
11. Lindsay RM, Ehrenberg ASC: The design of replicated studies. American Statistician 47:217–228, 1993
12. Berlin JA, Colditz GA: The role of meta-analysis in the regulatory process for foods, drugs, and devices. JAMA 281:830–834, 1999
13. Bridging Science and Service: A Report by the National Advisory Mental Health Council's Clinical Treatment and Services Research Workgroup. National Institute of Mental Health, 1998. Available at www.nimh.nih.gov/publicat/nimhbridge.pdf
14. Cochrane AL: Effectiveness and Efficacy: Random Reflections on Health Services. London, Nuffield Provincial Hospitals Trust, 1971
15. Brook RH, Lohr KN: Efficacy, effectiveness, variations, and quality: boundary-crossing research. Medical Care 23:710–722, 1985
16. Wells KB, Sturm R: Informing the policy process: from efficacy to effectiveness data on pharmacotherapy. Journal of Consulting and Clinical Psychology 64:638–645, 1996
17. Wells KB: Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. American Journal of Psychiatry 156:5–10, 1999
18. Fishman DB: Transcending the efficacy versus effectiveness research debate: proposal for a new, electronic "journal of pragmatic case studies." Prevention and Treatment 3, 2000 Available at www.journals.apa.org/prevention/volume3/pre0030008a.html
19. Norquist G, Lebowitz B, Hyman S: Expanding the frontier of treatment research. Prevention and Treatment 2, 1999. Available at http://journals.apa.org/prevention/volume2/pre0020001a.html
20. Stuart A, Ord JK: Kendall's Advanced Theory of Statistics, 5th ed: Vol 2. New York, Oxford University Press, 1987
21. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Edited by Stewart AL, Ware JE. Durham, NC, Duke University Press, 1992
22. Begg C, Cho M, Eastwood S, et al: Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA 276:637–639, 1996
23. Altman DG, Schulz KF, Moher D, et al: The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine 134:663–694, 2001