CHAPTER

3

Predictability: Greater than p < .05

Robert T. Wertz

The dread of appearing foolish is one of the biggest locks on the human cage. Therefore, those of us who conduct aphasia research usually do not report something unless we are 95 percent certain of it, for example, $p < .05$. However, even though we may have discovered a new "fact," we need to ask, "Who cares?" And, if we ask, we may find that not all facts are equal, and the best thing to do with some of them is to ignore them. We really want to know whether something applies always, sometimes, or never. We want to know how predictable our significant results are, because a powerful, predictable *fact* permits us to use the statistical properties of populations and make prejudgments about individuals from these populations.

Predictability permits us to distinguish between statistical significance and clinical significance. The former is precisely defined as the probability of something occurring by chance alone. And we place great emphasis on N, the number of individuals in a study cohort. Generally, we presume that the larger the N, the more valuable is the information obtained. However, the N can be so large that any association, even though of little or no clinical importance, can be made statistically significant. Investigators must clarify when the results of their research achieve statistical significance but not necessarily clinical significance. It is not helpful to be so in awe of the statistics that one fails to use informed judgment in translating the results of research into clinical practice. When statistical significance does not translate into clinical significance, the information obtained from research is misleading and is diminished as a resource for providing quality care. Further, it can lead to inappropriate clinical application; for example, the surgeon who read that 28 percent of all surgery is unnecessary, and therefore did only 72 percent of each operation.

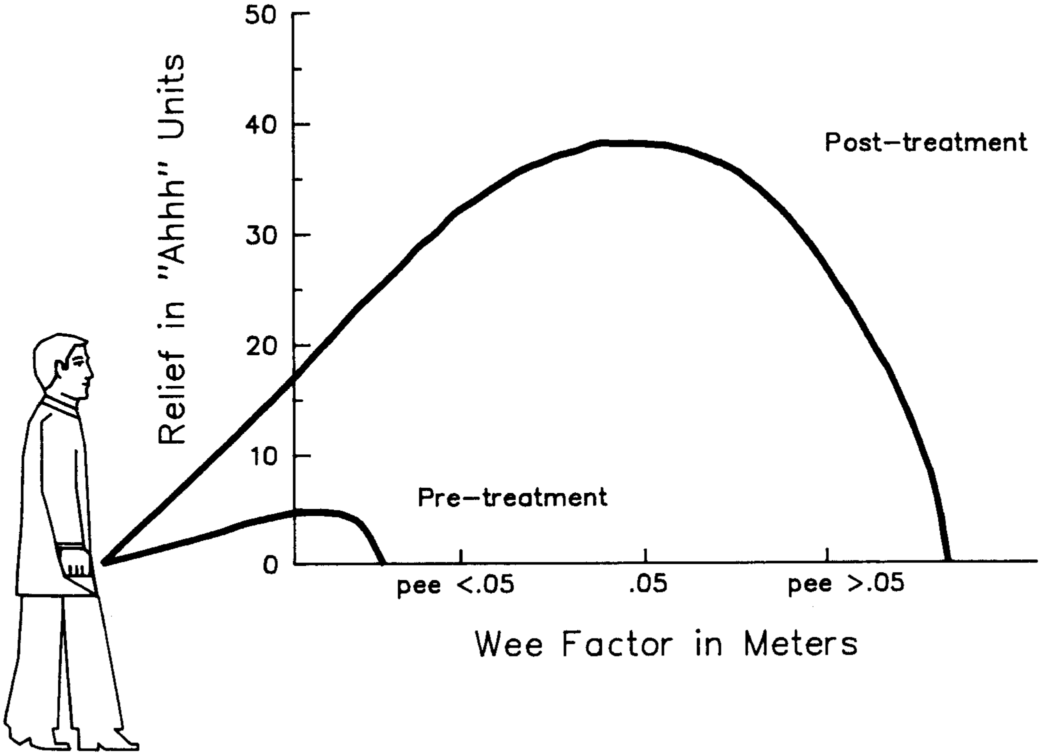
There are orthodox approaches to examining the predictability of one's results. These include calculating standard errors or confidence intervals or plotting receiver-operator characteristic (ROC) curves to determine how adjusting a cutoff affects sensitivity, specificity, and positive and negative predictive values. These approaches are elaborated by Longstreth, Koepsell, and van Belle (1987a, 1987b). Unfortunately, they are seldom found in the reports of results. I will not attempt to popularize them here except to imply that this is exactly what we need to be doing with our data to determine the significance of whatever significance we have found.

I have elected to take an unorthodox approach to discussing the predictability of research results and exploring the significance of significance. This approach, obviously, includes prostate dilatation, falling nightwatchmen, treatment for anthrax, Scholastic Aptitude Test gender differences, and the efficacy of treatment for aphasia. The data come from two sources: (1) a scientific review session of a proposal for surgical intervention in prostate dilatation and (2) an article by Sapolsky (1987). The purpose of what follows is to urge those who conduct aphasia research to ponder the significance of their significant results.

A STINT STUDY ON STUNTED STREAM

Our first lesson comes from the "Book of Urology." "In the land of prostate problems, it frequently comes not to pass, and there is no health in them." While there are effective methods for restoring urinary flow, the best, frequently, must be repeated. One of our urologists brought a research proposal to investigate the efficacy of a surgical stint in prostate dilatation before our scientific review committee. When questioned about his design, he explained that each patient would serve as his own control, and preoperative and postoperative performance would be compared. When asked about the variable he would be manipulating, he exploded, "What do you mean? What variable? These guys can't pee! We stint them, and they can." Thus the experimentalists on the com-

Fig. 3-1. Pretreatment and post-treatment comparison of surgical intervention in prostate dilatation. The results are powerfully predictable. Every patient's post-treatment performance is better than pretreatment performance.



mittee suggested a design that compared preoperative and postoperative performance on the "wee factor," measured in meters, plotted against relief, measured in "ahhh units." The early results are very positive and provide an excellent example of predictability.

As shown in Figure 3-1, preliminary results indicate that surgical intervention with a prostate stint provides a significant increase in both performance in terms of the "wee factor" and relief in "ahhh units." The separation of the curves demonstrates that every patient's posttreatment performance is improved over his pretreatment performance. And one can see a statistical rarity in the data. Pee greater than .05 is more desirable than pee less than .05. More important, the results are very predictable. Take any individual from the sample. If you know nothing more about him than whether he is preoperative or postoperative, and you prejudged his "wee factor" and relief in "ahhh units," you would be correct 100 percent of the time. This is a very powerful, predictable fact.

Now this paper flows, so to speak, in another direction, and the stream of inquiry asks, "When is a significant difference important, and when is it merely curious?"

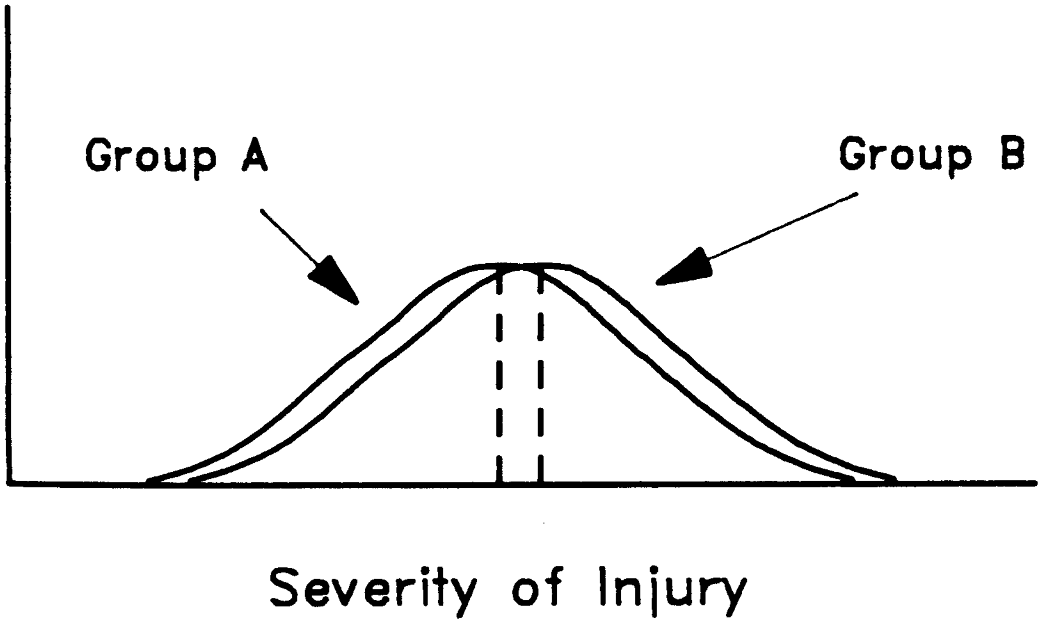
FALLING NIGHT WATCHMEN VERSUS TREATMENT FOR ANTHRAX

Our second lesson comes from the "Book of Sapolsky" (1987). He quotes the Alcoholics Anonymous prayer, "God, grant me the serenity to accept the things I cannot change, courage to change the things I can, and wisdom to know the difference." Sapolsky suggests that behavioral biology, and I include aphasiology in this ilk, is the scientific pursuit of that prayer. We ask which of our less commendable ways of behaving can we hope to change (and how) and which are we stuck with? Our research attempts to answer these questions, and the facts we discover to support our answers should be powerful and predictive, not merely curious.

Sapolsky uses analogy to differentiate between predictive facts and unpredictive facts. For example, he points out that the moon exerts a gravitational pull on the earth that changes over the course of a month as the moon's distance from the earth varies. Because human bodies are subject to the moon's gravitational pull, it follows that if night watchmen of equal weight and density tumble from the top of a building each midnight, the force with which they hit the ground will be influenced by the distance of the moon from the earth that night. An experiment could be designed to compare mortality and morbidity in group A of falling night watchmen, who tumble when the moon is close to the earth, with the severity of injury in group B of falling night watchmen, who fall when the moon is further from the earth. With sufficiently large samples, we could demonstrate, as shown in Figure 3-2, that the severity of injury in group A is significantly less ($p < .05$) than the severity in group B. However, if we attempted to predict that a given group B night watchman would be more severely injured than one from group A, Sapolsky observes we would be correct about 60 percent of the time. If we did not know the "fact" about the moon and guessed randomly, we would be correct 50 percent of the time. Knowing the "fact" is not much of an improvement. Certainly, we would not advocate a change in public policy to staff emergency rooms of hospitals more heavily when the moon's gravitational pull is less.

Conversely, Sapolsky reminds us that anthrax will kill you in about 48 hours. However, if antibiotics are administered immediately, you will recover. The survival curve, shown in Figure 3-3, for untreated anthrax patients, group A, does not overlap with the survival curve for treated anthrax patients, group B. Take an individual at random from each of the groups. If you knew nothing more about them than their group identities, and if you prejudged that the group B person would outlive

Fig. 3-2. Hypothetical falling night watchmen experiment in which group A watchmen, who fall when the moon is closer to the earth, sustain significantly less ($p < .05$) injury than group B watchmen, who fall when the moon is further from the earth. Results are significant but not very predictable. (Adapted from Sapolsky, 1987.)



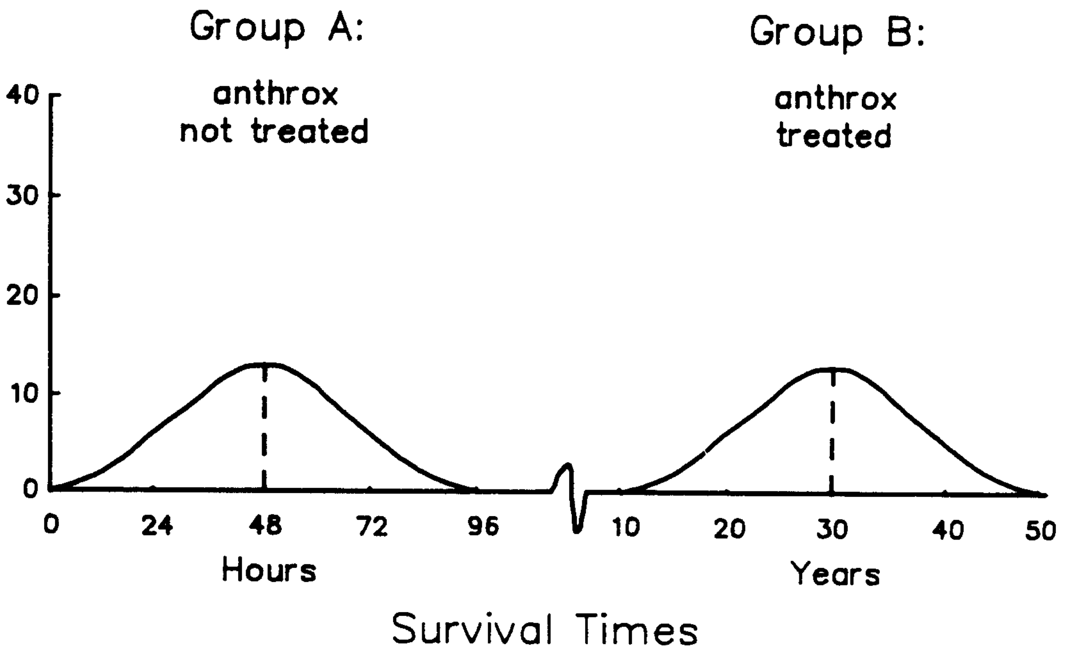


Fig. 3-3. Efficacy of treatment for anthrax indicates that all treated patients survive longer than all untreated patients. The results are powerfully predictable. (Adapted from Sapolsky, 1987.)

the group A person, you would be correct 100 percent of the time. If you didn't know their group identities, your chances of picking which person would live longer would be only fifty-fifty. The efficacy of treatment for anthrax is a powerful, predictive fact.

Finally, Sapolsky presents another "fact" that has surfaced repeatedly in our daily newspapers. Boys are smarter in math than girls. How do we know? Well, researchers have compared male and female scores on the mathematical section of the Scholastic Aptitude Test in 39,820 students and found that males achieved significantly higher scores than females ($p < .001$). However, being 99.9 percent certain does not ensure predictability. Sapolsky superimposed the performance curves, as shown in Figure 3-4, of males and females and noted that the average male score is 7 percent higher than that of females. He wanted to know what the predictiveness of this "factual" difference was. Take random pairs of boys and girls and—because you have read the research—prejudge that the boy in each pair will outscore the girl, and you will be right slightly less than 65 percent of the time. If you had no knowledge of the research and guessed randomly, you would have a 50 percent accuracy rate. Some improvement! Should we change our educational system and teach math differently to girls than we do to boys? Probably not. Should we discourage girls from developing their talents to the fullest in mathematics? Definitely not!

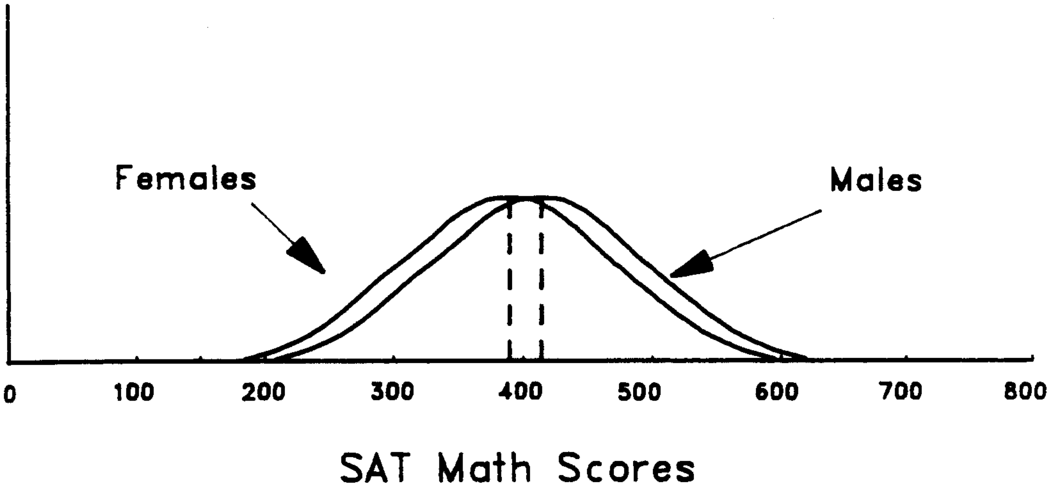


Fig. 3-4. Comparison of male and female performance on the math section of the Scholastic Aptitude Test. Males score significantly higher ($p < .001$) than females, but the results are not very predictable. (Adapted from Sapolsky, 1987.)

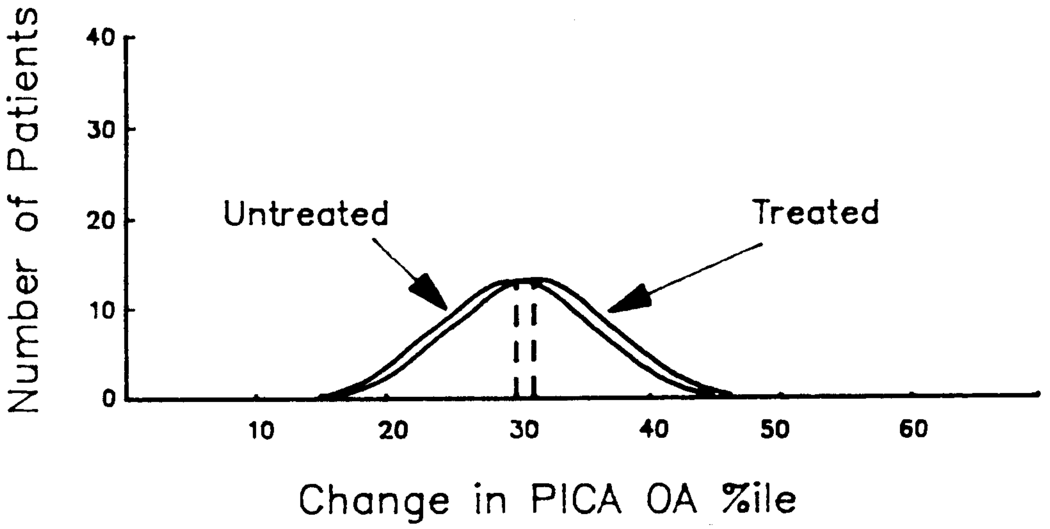


Fig. 3-5. Potential results in a study on the efficacy of treatment for aphasia. Treated patients make significantly more ($p < .05$) improvement than untreated patients, but the results are not very predictable.

PREDICTABILITY IN APHASIA RESEARCH

Finally, let us consider a fact whose predictiveness is yet to be demonstrated. We know that treated aphasic patients who meet specific selection criteria improve significantly more than aphasic patients who do not receive treatment (Wertz et al., 1986). How predictive is that little "fact"? It could be, as shown in Figure 3-5, that the overlap of the curves makes it impossible to predict that a particular treated patient improves

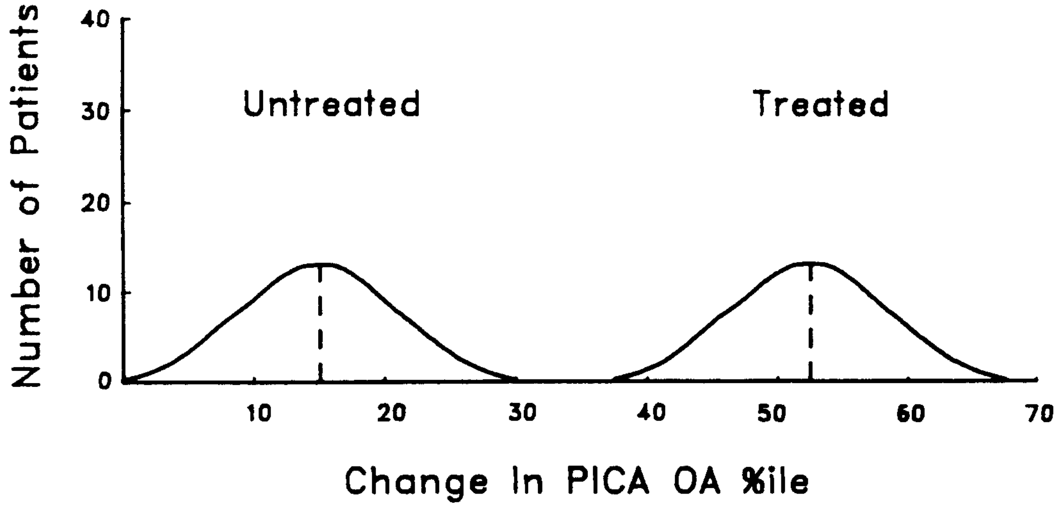


Fig. 3-6. Potential results in a study on the efficacy of treatment for aphasia. All treated patients make more improvement than all untreated patients. The results are powerfully predictable.

more than a particular untreated patient—essentially a falling night watchman or SAT result: significant, but who cares? Or it could be, as shown in Figure 3-6, that every treated aphasic patient improves more than every untreated aphasic patient—essentially a treatment for anthrax result: significant, and powerfully predictive.

Given the time, talent, and money expended on the treatment of aphasia, we really need to know whether the "fact" about the efficacy of treatment for aphasia is merely curious or powerfully predictive. Similarly, we would like to know whether the other significant "facts" we have established—the significant difference between aphasic and normal performance on the PICA (Duffy and Keith, 1980), the significant differences between language performance in aphasia and dementia (Bayles and Kaszniak, 1987), the efficacy of computerized language treatment for aphasia (Katz, 1987), and so on—are merely significant or powerfully predictive.

IMPLICATIONS

Anyone who has conducted research knows that a good deal of entropy occurs between the design and the data. I suspect a similar situation exists between our discovering a significant result, a "fact," and our interpretation of what that result means. You never want to let your brain get so full of facts that you have no room to think. Medawar (1984) observed that the equation of science with facts and the humane arts with ideas is one of the shabby genteelisms that bolster up the humanists' self-esteem. The scientific method is a potentiation of common sense exercised with a specially firm determination not to persist in error if any exertion of hand or mind can deliver us from it. Perhaps we turn off our minds too soon when we discover a scientific fact supported by statistical significance. Perhaps we should exercise our ideas by laying hands on our data and plotting their predictability.

From a metaphysical point of view, research in aphasia has not been linear. It has been cyclical. And the confusion and chaos in aphasia may indicate that we are approaching the end of not just a "cycle of time," but of a "cycle of cycles." When that end occurs, so much havoc and hockey will hit the fan that the whole fan will short circuit and the great Broca in the sky will be forced to play the greatest ace to keep all our efforts from biting the dust. Playing that ace may usher in the New Age of aphasia research. In the meantime, say the metaphysicians, if there is a name of god that is dear to you, keep it on your lips and you will be all right. Thus what we do makes metaphysical sense: By continuing to conduct research, we speed ourselves on toward the New Age, and by saying, "Good God! What do my results mean?" we keep a name of god on our lips.

Sapolsky (1987) concluded that

> Science is meant to make us free. It can give us wings with which to explore the world. And it can free us in a more mundane way: by making us aware of where there are walls we can't hope to push back. . . . Given these potential benefits, it's tragic when, by misreading the lessons of science, we're encouraged to leap off of cliffs when our wings are flimsy, or to feel constrained by walls that don't exist.

His cautions are sage. We must begin to ponder the significance of our significant results, provide some indication of their predictiveness, and determine whether our "facts" are important or mere curiosity.

REFERENCES

Bayles, K. A., and Kaszniak, A. W. (1987). *Communication and cognition in normal aging and dementia.* Austin, TX: PRO-ED.

Current Trends in the Practice of Medicine (1988). Pitfalls in statistical analysis. *Mayo Clinic Update, 4,* 6–7.

Duffy, J. R., and Keith, R. (1980). Performance of non-brain-injured adults on the PICA: Descriptive data and comparison to patients with aphasia. *Aphasia Apraxia Agnosia, 2,* 1–30.

Katz, R. C. (1987). Efficacy of aphasia treatment using microcomputers. *Aphasiology, 1,* 141–149.

Longstreth, W. T., Koepsell, T. D., and van Belle, G. (1987a). Clinical neuroepidemiology: I. Diagnosis. *Archives of Neurology, 44,* 1091–1099.

Longstreth, W. T., Koepsell, T. D., and van Belle, G. (1987b). Clinical neuroepidemiology: II. Outcomes. *Archives of Neurology, 44,* 1196–1202.

Medawar, P. B. (1984). *The limits of science.* New York: Harper & Row.

Sapolsky, R. M. (1987). Opinion: The case of the falling night watchmen. *Discover,* July, 42–45.

Wertz, R. T., Weiss, D. G., Aten, J. L., Brookshire, R. H., Garcia-Bunuel, L., Holland, A. L., Kurtzke, J. F., LaPointe, L. L., Milianti, F. J., Brannegan, R., Greenbaum, H., Marshall, R. C., Vogel, D., Carter, J., Barnes, N. S., and Goodman, R. (1986). Comparison of clinic, home, and deferred language treatment for aphasia: A Veterans Administration cooperative study. *Archives of Neurology, 43,* 653–658.