American Association of Electrodiagnostic Medicine

CLINICAL RESEARCH DESIGN IN ELECTRODIAGNOSTIC MEDICINE

Richard M. Dubinsky, MD, MPH
David S. Saperstein, MD
James M. Gilchrist, MD
Luca Padua, MD, PhD

2004 AAEM COURSE D
AAEM 51st Annual Scientific Meeting
Savannah, Georgia
Clinical Research Design
in Electrodiagnostic Medicine
Richard M. Dubinsky, MD, MPH

David S. Saperstein, MD
James M. Gilchrist, MD
Luca Padua, MD, PhD

2004 COURSE D
AAEM 51st Annual Scientific Meeting
Savannah, Georgia
Richard M. Dubinsky, MD, MPH  
Associate Professor  
Department of Neurology  
University of Kansas Medical Center  
Kansas City, Kansas  
Dr. Dubinsky earned his medical degree at the University of Missouri, and performed his internship and residency at Baylor College of Medicine in Houston, Texas. He then performed a fellowship in human motor control at the National Institutes of Health in Bethesda, Maryland. Dr. Dubinsky is currently an associate professor in neurology at the University of Kansas Medical Center, as well as the director of the electromyography laboratory there. He is also the associate director of the clinical neurophysiology fellowship program at the University of Kansas Medical Center and the workgroup chair of the "End of Life in Huntington's Disease" project. Dr. Dubinsky is a member of the Quality Standards Subcommittee of the American Academy of Neurology, and recently earned his MPH degree.  

David S. Saperstein, MD  
Assistant Professor  
Department of Neurology  
University of Kansas Medical Center  
Kansas City, Kansas  
Dr. Saperstein earned his medical degree from Boston University School of Medicine, and performed both his internal medicine and neurology residencies at Wilford Hall Medical Center. He is currently an assistant professor in the Department of Neurology, and in the Department of Pathology and Laboratory Medicine at the University of Kansas Medical Center. His academic interests include the clinical aspects of management of neuromuscular disorders with a special interest in the classification and identification of chronic acquired demyelinating polyneuropathies. In addition to membership in the AAEM, Dr. Saperstein is also a member of the Peripheral Nerve Society, the American Academy of Neurology, and the World Muscle Society.  

James M. Gilchrist, MD  
Senior Vice-Chair  
Department of Neurology  
Rhode Island Hospital  
Providence, Rhode Island  
Dr. Gilchrist earned his medical degree from Loyola-Stritch School of Medicine in Maywood, Illinois. He performed a residency and fellowship at the Medical College of Virginia, and then had a research fellowship in electromyography at Duke University Medical Center. He is currently Senior Vice-Chair in the Department of Neurology at Rhode Island Hospital in Providence, and both the director of the Division of EMG and Neuromuscular Disease and a professor in the Department of Clinical Sciences at Brown University School of Medicine. Dr. Gilchrist has been an ad hoc reviewer for several journals, including Neurology, Muscle & Nerve, Mayo Clinic Proceedings, and the Rhode Island Medical Journal. He is also on the Editorial Board for Muscle & Nerve.  

Luca Padua, MD, PhD  
Professor  
Department of Neurosciences  
Università Cattolica  
Rome, Italy  
Dr. Padua holds a medical degree with a specialization in neurology as well as a doctorate in neurosciences. He has been a researcher in the Department of Neuroscience and a professor of clinical neurophysiology at the Università Cattolica, as well as a professor of neurology. Dr. Padua's professional activities include positions as Coordinator of the Italian Carpal Tunnel Syndrome of the Italian Neurological Society, a member of the Advisory Board of the Italian Peripheral Neuropathy Study Group, and Coordinator of the Quality of Life Study Group of the Italian Neurological Society. He also serves as a member of the Editorial Board of the Journal of Sports Medicine and Physical Fitness, and a reviewer for Neurology, Muscle & Nerve, Neurological Sciences, and Clinical Neurophysiology.  

Authors had nothing to disclose.  

Course Chair: Richard M. Dubinsky, MD, MPH  
The ideas and opinions expressed in this publication are solely those of the specific authors and do not necessarily represent those of the AAEM.
Clinical Research Design in Electrodiagnostic Medicine

Contents

Faculty ii
Objectives iii
Course Committee iv

Introduction to Outcome Measures as Applied to Study Design 1
Richard M. Dubinsky, MD, MPH

The Use of Functional Outcome Measurements in Clinical Neuromuscular Disease Trials 11
David S. Saperstein, MD

Electrodiagnosis for Prognosis: Clinical Research Opportunities 17
James M. Gilchrist, MD

Development and Use of Quality-of-Life Tools 23
Luca Padua, MD, PhD

CME Self-Assessment Test 29
Evaluation 33
Future Meeting Recommendations 35

Please be aware that some of the medical devices or pharmaceuticals discussed in this handout may not be cleared by the FDA or cleared by the FDA for the specific use described by the authors and are “off-label” (i.e., a use not described on the product’s label). “Off-label” devices or pharmaceuticals may be used if, in the judgment of the treating physician, such use is medically indicated to treat a patient’s condition. Information regarding the FDA clearance status of a particular device or pharmaceutical may be obtained by reading the product’s package labeling, by contacting a sales representative or legal counsel of the manufacturer of the device or pharmaceutical, or by contacting the FDA at 1-800-638-2041.

Objectives
After attending this session, attendees will understand (1) the basics of the development and selection of outcome measures in clinical research, (2) the importance of well defined primary and secondary outcome measures, and (3) the necessity for the use of clinically relevant measures. Electrodiagnostic tests, functional measurements, and the development of quality-of-life instruments are the basic elements of this course. These measurements are applicable to diagnostic, therapeutic, and prognostic studies.

Prerequisite
This course is designed as an educational opportunity for residents, fellows, and practicing clinical EDX consultants at an early point in their career, or for more senior EDX practitioners who are seeking a pragmatic review of basic clinical and EDX principles. It is open only to persons with an MD, DO, DVM, DDS, or foreign equivalent degree.

Accreditation Statement
The AAEM is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education (CME) for physicians.

CME Credit
The AAEM designates attendance at this course for a maximum of 3.5 hours in category 1 credit towards the AMA Physician’s Recognition Award. This educational event is approved as an Accredited Group Learning Activity under Section 1 of the Framework of Continuing Professional Development (CPD) options for the Maintenance of Certification Program of the Royal College of Physicians and Surgeons of Canada. Each physician should claim only those hours of credit he/she actually spent in the activity. The American Medical Association has determined that non-US licensed physicians who participate in this CME activity are eligible for AMA PMR category 1 credit.
2003-2004 AAEM COURSE COMMITTEE

Kathleen D. Kennelly, MD, PhD
Jacksonville, Florida

Thomas Hyatt Brannagan, III, MD
New York, New York
Kimberly S. Kenton, MD
Maywood, Illinois

Dale J. Lange, MD
New York, New York
Andrew Mazur, MD
Portsmouth, Rhode Island
Christopher J. Standaert, MD
Seattle, Washington

T. Darrell Thomas, MD
Knoxville, Tennessee
Bryan Tsao, MD
Shaker Heights, Ohio

2003-2004 AAEM PRESIDENT

Lois Margaret Nora, MD, JD
Rootstown, Ohio
Introduction to Outcome Measures as Applied to Study Design

Richard M. Dubinsky, MD, MPH
Department of Neurology
University of Kansas Medical Center
Kansas City, Kansas

INTRODUCTION

Medical journals serve to share information about the diagnosis and treatment of conditions. They then generalize these results to the diagnosis and treatment of other patients. Given the size of the human population, it is not possible to completely describe the entire population, to understand all diseases, and identify all possible contributions to disease and health. With these profound limitations, physicians must rely upon their interpretation of the medical literature and their ability to perform well-executed randomized clinical trials (RCTs) to answer clinical questions.

Over time the medical literature has moved from case reports to RCTs. First in the treatment of infectious diseases and then in the treatment of cancer, RCTs were adopted in the United States in the late 1940s and 1950s. Many of the designs currently employed in RCTs were developed at the National Cancer Institute (NCI), including the multi-center treatment trial design. The objectivity gained by the use of controlled trials has raised the standards for medical research. Study design in clinical neurophysiology has progressed little from the 1940s and 1950s and has fallen behind the advances in RCTs in other fields of medicine. For the science of clinical neurophysiology to mature with the rest of the clinical sciences, physicians must improve the design and execution of their RCTs. The objective of this manuscript is to introduce the basic concept of RCTs to clinical neurophysiology, concentrating on the selection and design of appropriate outcome assessment measures.

The first RCTs in the United States were sponsored by the Department of Agriculture in the 1930s and 1940s. The yield of different strains of grains was tested through random assignment to farmers’ fields with the results of the harvest measured in an objective fashion. In the 1950s the NCI introduced RCTs in the evaluation of chemotherapy for leukemia. In these trials, subjects were randomized to receive either standard chemotherapy, or standard chemotherapy plus the experimental therapy. The outcome measurement was simple—survival. This outcome was unambiguous and easy to determine. Binary outcomes are the easiest of outcome measures to work with, but are rare to non-existent in clinical neurophysiology. Unfortunately, most outcome measures are not as clear-cut and are subject to bias in their interpretation and reporting.

Case reports and small case series are used to report the results of open-label trials of innovative treatments. The authors have a vested interest in the outcomes that they are determining. This leads to bias, or prejudice, in the reporting of the results towards benefit from the treatment. A recent meta-analysis compared the rates of benefit reported between open-label (nonrandomized) therapeutic trials and RCTs of the same treatments. The meta-analysis utilized odds ratios to compare the results of the open label studies to those from the RCTs. Of the 45 pairs of studies, one-third (15) had a greater than or equal to two-fold inflation of benefit comparing the open-label to the RCTs, and almost two thirds (28) had a greater than or equal to 50% inflation of benefit. This trend was similar across the types of studies, including prospective, retrospective with current controls (case control), and studies with historical controls. A similar reduction in the
activated, recombinant tissue plasminogen (rTPA) is approved by the Food and Drug Administration for treatment of myocardial infarction and acute cerebrovascular accident. In the initial National Institute of Health (NIH)-funded study, rTPA increased the proportion of stroke survivors with a moderate improvement after treatment, when compared to the placebo. This improvement was found not at 24 hours after treatment, but rather at 3 months after the stroke. The rate of symptomatic intracerebral hemorrhage at 36 hours was greater for the rTPA group (6.4% vs. 0.6%) and the mortality rate at 3 months was less (17% vs. 21%). Community results are quite different. In several studies, the mortality was reported to be between 9.9% and 15%, compared to the initial mortality rate in the NIH funded study. Some of the excess mortality was determined to be due to protocol violations, treating patients who would not have been treated in the original study. In this author’s own analysis of the Nationwide Inpatient Sample, a 20% stratified sample of all hospitalizations in the United States of in-hospital mortality from ischemic strokes for the years 1999 through 2002 increased from 6.4% to 12.8% when patients were treated with rTPA. The rate of intracerebral hemorrhage went from 0.6% to 3.6%. Some of the excess mortality may be due to differences in stroke severity or comorbid conditions, but still the excess mortality is worrisome. One way of looking at this is the disclaimer often found on infomercials: Individual results may vary.

Outcome measures are not always dichotomous. Most often they are continuous variables that physicians try to transform into ranges of normal, abnormal, or in between. For example the normal serum cholesterol is often defined as less than 210 mm/dL. However, one-third of patients with their first myocardial infarction have a serum cholesterol between 150 and 210 mg/dL. Using the accepted definitions applied to cardiac risk, there is an overlap between the normal and diseased ranges. Before the determination of outcome measures can be discussed, some basic concepts of biostatistics need to be established.

**BIOSTATISTICS**

Statistics is the science of making decisions based upon incomplete data or information. The results can be described in terms of p or probability or in terms of odds. Probability is the chance of something that one wants to occur against all the possible outcomes. The probability of rolling a 3 from a roll of a die is 1 in 6 while the probability of rolling an even number is 3 in 6 or 1 out of 2 rolls. Odds are defined as the chance of what one wants to occur compared to the other possible results. Thus the odds of getting a 3 from a roll of a die is 1 in 5 and the odds of rolling an even number is 3 in 3, or even odds. The odds are the same for rolling an even or an odd. When the number of possible results increase, the odds of the desired response approaches the probability. The odds of winning the multi-state lottery are 1 in 120,526,770 while the probability is 1 in 120,526,771. Both are close, but not quite zero.

In biostatistics, decisions are made based upon incomplete knowledge of biological systems. A population can never fully be described, only samples of the population. Thus, the median motor distal latency of a group of adults who meet the clinical case definition of carpal tunnel syndrome (CTS) can be determined and compared to the results of those who are asymptomatic. The spread of measurements in a biological population, expressed using a histogram, is almost always a bell shaped curve or a Gaussian distribution. This is also called a parametric or normal distribution. The mean (average response) is easy to determine, as is the median (or middle) response. If a sample is large enough, the mean and median will be quite close. If the sample is small, the distribution will be skewed and the mean and median will be farther apart. In large samples the distribution can be skewed when the distribution is not parametric. Another way of describing the distribution is to combine the mean and the standard deviation (SD) (or spread) of the sample. The mean ± on 1 SD describes two-thirds of the responses. The mean ± 2 SDs, describes 95% of the sample, with 2.5% of the results below and 2.5% above. The important thing to remember about the use of a mean ± 2 SDs of a population is that 2.5% of the population is below and 2.5% is above the limits of the 2 SDs. To state this again, 1 in 20 members of a normal population will be beyond the boundaries used to define the normal population.

When a population is not under a normal distribution it is described by the use of the median (middle) response and the range (minimum to maximum) of the responses. This is a more inclusive description than what is used for a normal population. Since parametric and nonparametric populations have different distributions, different mechanisms are needed to compare them. Parametric tests, like the Students t test are used to compare the mean of two populations while the nonparametric tests, like the Wilcoxon signed rank test are used to compare the central tendencies of two populations. The problem when comparing two distributions of a population (such as our example of the median motor distal latency in a normal population and a population that meets the clinical case definition of carpal tunnel syndrome) is that they can overlap.

So far what has been discussed applies to only one type of data: continuous. Most clinical neurophysiology results are continuous numbers with the problem being the definition of the normal and abnormal ranges. However, other types of data are involved in describing the populations. These other types of data include dichotomous, nominal (naming), categorical, and
ordinal (named ordering). Dichotomous data has only one of two values, such as survival and death. Nominal data is named or labeled data, with no inherent ordering. The easiest application of nominal data is the use of control and experimental groups. Categorical data are one application of nominal data, such as ethnicity, race, and occupation types. Ordinal data are categorical with inherent ordering. Grades in school and belt mastery in martial arts are two examples of this type of data. The most important thing to remember about ordinal data is that they are still categorical data, and not numerical data.

A common mistake is to try to average ordinal data. There is no inherent meaning in reporting that the average educational level of a population was 12.5 years. A common application of this type of mistake is the reporting of the results of a Likert, or numerical scale. Hospital administrators frequently use a numerical scale to judge physician satisfaction and then report the results as “we have improved physician satisfaction from an average of 1.7 to 2.3 out of a possible 5.” The administrators missed two important points. One is that they should report the proportion of physicians who felt very dissatisfied (1), dissatisfied (2), neutral (3), satisfied (4), and very satisfied (5). The second is that the trend of the physician responses is towards dissatisfaction.

Statistical tests are used to compare two or more populations. The result of the comparison is expressed as the probability (p) of the same results occurring by random chance. Thus a p = .05 means that if the experiment was repeated 20 times in subsets of the same population, 1 of the 20 times the difference between the two populations would be the same or more extreme, due to random error. This also means that if 20 tests are performed on a population, 1 of the trials will show a significant difference due to random error. This leads to the next several topics: determining primary outcome assessment measures, the degree of difference to be measured, how many subjects are required to find that difference, and data mining.

**HYPOTHESIS TESTING**

The first step is to establish the null hypothesis (Ho) that there is no difference in the measure between the two populations. The alternate hypothesis (Ha or H1) is that there is a difference in the measure between the two populations. The statistical tests employed serve to reject the null hypothesis and to prove the alternate hypothesis. One can never prove the null hypothesis, one can only fail to reject it. These hypotheses need to be stated in terms of the primary outcome measurement.

- Ho: There is no difference in the median motor distal latencies between the two groups.
- Ha: There is a difference of > 0.5 ms in the median motor distal latencies between the two groups.

The secondary outcome measures can be included into additional sets of null and alternate hypotheses, but these should not be examined statistically until (and only if) the primary outcome measure shows a significant difference.

**POWER CALCULATION**

Study trials are designed to have a reasonable chance of finding a significant difference, provided that difference exists. This is called the “power” of a study. Power is calculated based on the estimated difference between the study groups (e.g., placebo vs. active drug), the likelihood of finding the difference (beta, typically set at 80% or 90%), the probability of finding a difference based on random error (alpha, usually set at .05 or less) and the number of subjects (n). A power calculation is read as stating if n subjects are studied, there is an 80% chance of finding the specified difference between the two groups, at a significance level of p = .05, if the difference exists. Usually a power calculation is used to determine the number of subjects needed in each group. When there are a limited number of subjects, as in the case of a rare condition in a restricted population, power can be calculated based on the number of subjects.

In an RCT, the number of subjects needed to answer the clinical question is based upon the magnitude of benefit estimated from prior studies; however, the number of confounding variables that could influence the outcome of the trial must be taken into consideration. A confounder is any variable that could potentially account for a difference between the two groups. Typical confounders are age, gender, and health conditions other than the one under study. In a study of treatment of CTS, a group with a greater proportion of older subjects, those with diabetes, or those with occupations requiring excessive hand movements, may be expected to have a poorer outcome. With more confounding variables, more subjects are needed to determine the outcome, which increases the cost of the study. The comorbidities in the population in question may have been excluded in the original trial. Or, in the case of rTPA, the treatment may not be given within the optimal 180-minute window. Thus the tight inclusion and exclusion criteria used in the RCTs are not always followed in community use of a treatment. This can decrease the magnitude of benefit, decrease the chance of benefit, and increase the chance of harm. This further limits the ability to generalize...
the results of treatment in a restricted study population to any given patient.

**DATA MINING**

If enough tests are performed on different variables between two populations, eventually a significant difference will be found due to random error. Because of this fact, it is important to determine the desired level of significance before the study begins and to then adjust this if multiple comparisons will be performed. Otherwise, there is a risk of data mining—searching for an outcome measure with a significant difference between the two populations and then publishing only that variable while ignoring all of the outcome measures that did not have a significant difference.

To understand research, it is important to define the terms used as well as to discuss bias.

**Case definition:** This is the basic definition of a disease state. Occasionally this can be determined with the use of a laboratory test or a pathology specimen. More often than not it is a set of statements that, if the patient meets them, causes the researcher to believe the patient has the condition. A consensus care definition was developed in 1998 for the purpose of epidemiological studies of CTS. These definitions reflected the likelihood of CTS based on symptoms (Table 1).

<table>
<thead>
<tr>
<th>Classic probable</th>
<th>Numbness, tingling, burning, or pain in at least two of digits I, II, or III. Palm pain, wrist pain, or radiation proximal to the wrist is allowed.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possible</td>
<td>Tingling, numbness, burning, or pain in at least one of digits I, II, or III.</td>
</tr>
<tr>
<td>Unlikely</td>
<td>No symptoms in digits I, II, or III.</td>
</tr>
</tbody>
</table>

This is quite different from the electrophysiological criteria (prolonged median distal sensory latency, prolonged median motor distal latency) proposed for the diagnosis of CTS.

**Experimental Group:** The group of subjects who receive the experimental treatment or test.

**Control Group:** The group of subjects who receive the conventional treatment or test.

**Disease Population:** The overall population with a disease. The experimental and control groups are part of this population. This may include those at risk for a disorder, or it may include those with the same complaints or physical findings due to a different disease. In that case a diagnostic test may be evaluated to determine not if it can be used to differentiate disease from normal, but a specific disease (CTS) in comparison to a similar disorder (lower cervical radiculopathy).

**Population Universe:** The entire population. This can never be completely examined with humans, but it can be described through sampling.

**Incidence:** The number of new cases of a disorder appearing in a population over a defined time period.

**Prevalence:** The number of cases in a population during a defined time period. This is a function of incidence and survival.

**THE THORNY ISSUE OF BIAS**

Bias is anything that introduces a nonrandom change in the results of a study. Thus bias is anything that can move the results away from being purely objective and reproducible. Some forms of bias are obvious; many are not. Dr. David Sackett, one of the early proponents of evidence-based medicine, described over 30 types of bias that can occur in medical research. A few of the more obvious, and not so obvious forms of bias will be discussed.

**RECRUITMENT BIAS TOWARDS HEALTHY SUBJECTS**

Clinical trials can cost an inordinate amount of money. The more heterogeneous the trial, the more confounders need to be considered, and the more subjects are needed. If the subjects are homogeneous, it is easier to determine a treatment effect, and there are fewer confounders to control. Getting the healthiest of the sick into a treatment trial also can inflate the benefit of the treatment and prevent the study from being generalized to other populations. Other types of this form of bias include recruiting only one gender, a narrow spectrum of racial and ethnic groups, a narrow age range, or a narrow selection based on employment or educational level when there is no biological hypothesis for this narrow selection. This also prevents the study from being generalized to other populations. However, if the purpose of the study is to detect differences in median sensory nerve responses
in women who meet the clinical case definition of CTS, including women during their third trimester of pregnancy, inclusion of males, or nonpregnant women who use their hands in manufacturing, the outcome is nonsensical.

**EASY TO DISTINGUISH BETWEEN DISEASE AND NORMAL SUBJECTS**

Applying a clinical test between those who are known to have a disease (through the use of a reference standard) and a normal population yields easily interpreted results. It can be argued that if a subject does not meet the clinical case definition of CTS, and has no complaints of hand numbness, weakness or pain, that to perform electrodiagnostic (EDX) studies to determine if they have CTS is pointless. It is easy to distinguish health versus disease. What needs to be tested is the utility of a test to distinguish two similar disease states (e.g., CTS and cervical radiculopathy).

**LACK OF OBJECTIVE ASSESSMENT**

In any study the experimenter hopes to detect a difference and in treatment trials the subjects expect to feel benefit. Certain outcomes are objective, such as survival versus death, the number of myocardial infarctions, and the number of recurrent strokes. Others, such as the sensory examination and global impression of change are not objective. If the person who makes the ultimate assessment has an inherent interest in the outcome of the study (a surgeon reporting a new operative technique for low back pain) he or she may intentionally or unintentionally bias the results of the study towards favoring the new treatment or test. This is no different from a fund manager touting the success of their investment fund or the accuracy of their accounting methodologies.

Blinding of the investigator and patient to treatment assignment can mitigate this type of bias in the study of therapies. Blinding of the investigator to the disease status, when interpreting radiology studies or waveforms is another method to eliminate this type of bias. For more on blinding, see the attached manuscript from the AAEM, “The Need For Blinded Electrodiagnostic Studies” (Attachment A).

**RANDOMIZATION FAILURE**

Studies of therapies should have random assignment of subjects to the treatment groups. If the assignment is nonrandom, then the investigator can assign healthier patients to one treatment group, or those with more clear-cut physical findings to a specific diagnostic test group. Many seemingly random forms of cohort assignment are easily subverted. These include assignment by day of the week that the subject was recruited, year or date of birth, odd or even day, or by position of the first letter of their name in the alphabet. None of these provide true random assignment.

**TREATMENT ASSIGNMENT FAILURES (INTENT TO TREAT)**

When subjects are assigned to a treatment group, they do not always stay there. They may seek treatment outside of the study, they may be mistakenly given the nonassigned treatment, or they may drop out before receiving treatment. In a recent study from the Netherlands, the outcome of CTS was compared between an early surgery group and a conservative treatment group. At the end of the 18-month-long study there was only a 15% difference in the rate of symptom-free subjects between the two groups. However, 33 of the 79 subjects assigned to the conservative therapy group switched treatment assignment and underwent carpal tunnel surgery during the course of the study.

**DROP OUT AND LOSS TO FOLLOW-UP**

Not all subjects who start a study will finish the study. In some of the early studies of centrally acting acetylcholinesterase inhibitors for Alzheimer's disease, over 60% of the subjects dropped out of the study due to liver toxicity while on an active drug. For those who could tolerate the treatment, there was a mild but significant benefit. When the majority of subjects who could not tolerate the treatment were included in the assessment, the significance was lost. As an aside, their primary outcome assessment was the Mini-Mental State Examination (MMSE) a screening tool used to detect cognitive impairment. A difference in the rate of decline of the MMSE by 0.7 points per year is clinically meaningless.

**CONCEALMENT OF ADVERSE OUTCOMES**

Concealment of adverse outcomes has been in the national media in the United States frequently over the last month in regards to clinical trials of medications, in particular the selective serotonin reuptake inhibitor family of antidepressants. There is now a strong push to have RCTs registered before their inception, and to have the database of subject results available for public and scientific review. The reasoning behind these steps is that the FDA and others believe that not all adverse events are being reported leading to a false sense of safety regarding several treatments. Certainly if a surgeon investigating the results of carotid endarterectomy surgery fails to disclose a post-operative
stroke, because the nonoperated carotid artery was involved, the danger of the surgery would be minimized.

**PUBLICATION**

One recent method adopted to improve the quality of published scientific studies has been the series of statements by the Consolidated Standards of Reporting Trials group. The work group established over 20 items that should be included in any clinical trial. These include cohort assembly, methods section sufficient to allow reproduction of the study, a flow diagram showing the flow and narrowing of subjects through the study, a sample of the raw data, and tabular summary of the entire dataset. The cohort statement and the new clinical trials registries are attempts by the medical community to raise the standards for RCTs. Most, if not all, of the pioneering studies in clinical EDX medicine met the scientific rigor of their times. There is no need to repeat these studies using modern clinical trial methodologies. But, there is a strong need for current and future studies in EDX medicine to embrace the higher level of scientific rigor that has been adopted in clinical medical research. Failure to improve the quality of clinical neurophysiology scientific studies will only lead to being left behind in the dust of techniques in neuromuscular medicine developed by those outside of our field. The rules have changed. The new rules need to be adopted and the scientific rigor of the clinical neurophysiology field improved.

---

**BIBLIOGRAPHY**

Abstract: The level of scientific rigor has advanced since many of the electrodiagnostic (EDX) tests were developed. Currently, there is a paucity of blinded studies (i.e., where the person evaluating the data does not know the clinical details) in the research assessment of EDX tests. Lack of blinding is now recognized as a major source for the introduction of bias into the performance and interpretation of diagnostic tests in the research arena. The AAEM recommends that a standard of blinding of the investigator or independent assessment be used in future research of EDX techniques as applied to diagnosis, prognosis, and outcomes research.

INTRODUCTION

The American Association of Electrodiagnostic Medicine’s mission is to serve physicians who diagnose and treat patients with disorders of muscle and nerve, extend the knowledge of electrodiagnostic medicine, and improve the quality of patient care. Improving the quality and credibility of research in the field of electrodiagnostic medicine (EDX) will help accomplish this mission. Currently, there is a paucity of blinded studies (i.e., where the person evaluating the data does not know the clinical details) in the research assessment of EDX tests. To meet the current research standards, studies of the accuracy of EDX techniques must be either blinded or incorporate independent assessment. The AAEM Board of Directors therefore charged a task force to draft this document outlining the need for blinded research and providing recommendations regarding how to perform blinded EDX research in the future. The AAEM’s goal is to facilitate and encourage more blinded EDX research.

BACKGROUND

Clinically, EDX studies are used for diagnosis and for prognosis. The clinical EDX consultation is based upon the clinical interview, the examination, and the EDX evaluation. The utility of EDX testing in clinical practice is based upon the findings of the usefulness of the technique when studied in a controlled research environment. In the research setting, the EDX evaluation is examined in isolation from the clinical setting.

The scientific standards for clinical research have advanced since the important fundamentals of electrodiagnosis were established. Scientific standards now embrace items such as human subject concerns, bias issues, reproducibility with clear presentation of methodology, inclusion of samples of the raw data as part of the results, accountability to sponsoring organizations, publication standards, and the need to eliminate scientific fraud. The impact of the earlier studies, which do not necessarily meet these new standards, is not lessened. However, now it is important that the scientific rigor of future clinical studies be increased. One way to increase the scientific rigor is by minimizing bias, which is the introduction of systematic error into a study.

Bias can affect the selection of the study population, performance of diagnostic tests, and interpretation of test results. Controlled clinical trials are an important part of eliminating bias. In nonrandomized clinical trials, there is often inflation of the magnitude of benefit. A study by Ioannidis and colleagues has shown that nonrandomized studies frequently overestimate the magnitude of benefit by 40% when compared to randomized studies.¹

Bias is obvious in the report of a therapeutic intervention where both the intervention and the determination of benefit or harm are reported by the same person in an unblinded fashion. Other examples of bias are when a researcher has a preconceived notion of the proper outcome, a personal/economic interest in the
outcome, or is motivated to publish favorable results. All of these
bias the conduct of the study towards a particular outcome.
Blinding is one method to overcome bias.

Blinding, while accepted in many disciplines, is not used in most
current research publications of EDX tests. Of the 103 abstracts
presented at the 2002 AAEM Annual Scientific Meeting, only
four included any form of blinding. Of the four studies, one
described the sensitivity of a test in a patient population (no
control group) while the second described the prognosis after
surgical treatment. The third used EDX as an outcome measure
in a therapeutic trial and the fourth stated that an independent
blinded review was also being performed, but no such data was
provided in the abstract. Of the remaining abstracts, 27 were
unblinded studies of EDX techniques in patient populations, 20
were of diagnostic tests in normal populations, 44 were case
series (mostly retrospective), 3 were basic science research, 2
were psychological testing of instrument displays, 1 was a survey
of clinicians, and in 2 cases the study design could not be deter-
dined. Clearly there is a need for EDX physicians to understand
the importance of blinding as applied to EDX research.

EVALUATION OF DIAGNOSTIC AND PROGNOSTIC TESTS

The AAEM in recent years has worked closely with the American
Academy of Neurology (AAN) to develop practice parameters
that assess electrodiagnostic tests in the diagnosis of ulnar neu-
ropathy and carpal tunnel syndrome (CTS). Table 1 classifies
the current criteria used in the systematic evaluation of diagnos-
tic tests for the development of practice parameters as used by
the AAEM and AAN. These classes of evidence present an
ordinal scale for studies based on the potential for bias.

The issue of whether an independent, blinded comparison, with
a reference standard (formerly known as the gold standard) was
utilized when conducting the study is a critical component of
any evaluation of a diagnostic tool. Just as one may question the
objectivity of a study to determine the accuracy of magnetic
resonance imaging (MRI) in the diagnosis of radiculopathy
where the radiologist performs the MRI and the myelogram
(reference standard), current research results are questioned
when the assessment is either unblinded or not independent of
the outcome. This is similar to the need for a placebo control in
treatment studies.

<table>
<thead>
<tr>
<th>Class</th>
<th>Evidence provided by a prospective study in a broad spectrum of persons with the suspected condition, using a reference standard for case definition, where test is applied in a blinded evaluation, and enabling the assessment of appropriate tests of diagnostic accuracy.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class II</td>
<td>Evidence provided by a prospective study of a narrow spectrum of persons with the suspected condition, or a well designed retrospective study of a broad spectrum of persons with an established condition (by reference standard) compared to a broad spectrum of controls, where test is applied in a blinded evaluation, and enabling the assessment of appropriate tests of diagnostic accuracy.</td>
</tr>
<tr>
<td>Class III</td>
<td>Evidence provided by a retrospective study where either persons with the established condition or controls are of a narrow spectrum, and where test is applied in a blinded evaluation.</td>
</tr>
<tr>
<td>Class IV</td>
<td>Any design where test is not applied in blinded evaluation OR evidence provided by expert opinion alone or in descriptive case series (without controls).</td>
</tr>
</tbody>
</table>

*Current standards as developed by the Quality Standards Subcommittee, and the Therapeutic and Technology Assessment committees of the American Academy of Neurology. Emphasis added by the authors.

Under the current scientific standards, the diagnostic test must
be compared to a reference standard. The standard may be based
on pathological material, imaging studies, established EDX
techniques, or a case definition. This comparison is necessary if
the goal is to prove that the new technique is as accurate or more
accurate than the reference standard. For many disorders, the
case definition may be the only reference standard. For example,
in CTS the current case definition is based on historical and
examination features. The use of clinical criteria in the case
definition of CTS, rather than the EDX criteria, would prevent
the misdiagnosis of median mononeuropathy in an asymptomatic
worker with prolonged median sensory nerve action potential
latencies.

Biases of interpretation can occur when evaluating a diagnostic
or prognostic test. These biases are overcome with the use of

Table 1 Classification of the Level of Evidence of Diagnostic Studies

<table>
<thead>
<tr>
<th>Class</th>
<th>Evidence provided by a prospective study in a broad spectrum of persons with the suspected condition, using a reference standard for case definition, where test is applied in a blinded evaluation, and enabling the assessment of appropriate tests of diagnostic accuracy.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class II</td>
<td>Evidence provided by a prospective study of a narrow spectrum of persons with the suspected condition, or a well designed retrospective study of a broad spectrum of persons with an established condition (by reference standard) compared to a broad spectrum of controls, where test is applied in a blinded evaluation, and enabling the assessment of appropriate tests of diagnostic accuracy.</td>
</tr>
<tr>
<td>Class III</td>
<td>Evidence provided by a retrospective study where either persons with the established condition or controls are of a narrow spectrum, and where test is applied in a blinded evaluation.</td>
</tr>
<tr>
<td>Class IV</td>
<td>Any design where test is not applied in blinded evaluation OR evidence provided by expert opinion alone or in descriptive case series (without controls).</td>
</tr>
</tbody>
</table>

*Current standards as developed by the Quality Standards Subcommittee, and the Therapeutic and Technology Assessment committees of the American Academy of Neurology. Emphasis added by the authors.
blinded assessment and comparison to a reference standard. In studies that rely on radiographic findings for outcome assessment, the investigator is blinded to the clinical status and to the results of any tests performed with the reference standard. When the studies are blindly presented to the investigator from a study population that is a mixture of normal and diseased subjects, accurate assessment of the test is possible. The addition of clinical information has the potential to bias the research investigator’s interpretation of the test. Studying a homogeneous population (either all normal or all diseased) precludes the assessment of diagnostic accuracy. A heterogeneous population (including normal subjects, other disease controls, and the disease under investigation) must be studied in order to determine the sensitivity (how many of those with the disease have an abnormal result) and specificity (how many who are disease-free have a normal result).

HOW TO PERFORM A BLINDED EVALUATION OF AN ELECTRODIAGNOSTIC TEST

Nerve conduction studies (NCSs) are considered objective, reproducible tests, however, since there are subjective components associated with NCSs it could be subject to bias. The take-off and peak latency are frequently marked manually and could be subject to bias. The effort expended can also be subject to bias when the result is difficult to obtain. Consider, for example, the evaluation of CTS. If the latency difference is borderline, there might be a tendency to change the marker to minimize or maximize the difference based on the knowledge of the other results. If the response is of small amplitude, it may be interpreted as not present if the researcher knows that the reference standard test result was abnormal. The needle examination is more prone to subjective assessment.

NCSs are reasonably reproducible but still are subject to bias. A simple mechanism to evaluate the bias would be to have a second clinician mark the take-off and peak latencies on the waveforms without knowledge of the clinical history and examination. This can be accomplished very easily with current computerized machines. This evaluation is considered single blind, if the researcher does not have access to the clinical information. It is double blind if the subject did not yet possess a diagnosis based on the reference standard when the waveforms were collected. Alternately, a blinded researcher could independently perform the entire study.

In a similar fashion the needle examination, which most EDX consultants would agree is more subjective, could be evaluated by an independent clinician without knowledge of the clinical history or physical examination. For example, a second EDX investigator could be involved while the study is taking place. This researcher would not know the clinical history or examination results but would be in the examination room while the needle examination is taking place. Ideally, the second clinician would not see the patient since the observation of muscle atrophy or functional limitation could create potential bias. The second observer would know only which muscles were being tested and determine the findings independently of the clinician performing the study. The second evaluator would not direct the study or provide information to the clinician. The independent evaluator would reach a conclusion that then could be compared to the reference standard or case definition.

RECOMMENDATIONS

The AAEM recommends that a standard of blinding of the investigator or independent assessment be used in future studies of EDX techniques as applied to diagnosis, prognosis, and outcomes research.

There must be blinded assessment or independent assessment wherever possible. This requires the separation of the roles and the information available to the clinician and the researcher. The clinician must have all of the needed information to make a determination and to insure that patient care is not compromised. The researcher must have access only to the data that is necessary for performance and evaluation of the research test.

The extent of blinding in an individual study may vary with the research question. EDX consultation is an interactive process involving performance of individual tests, choice of an array of individual tests, interpretation of the findings in isolation, and interpretation of the findings in the context of a history and physical examination. Studies examining each of these phenomena may require different levels of blinding. For example, a study of the sural sensory nerve conduction response might be appropriately blinded by having an examiner simply read a printout of a study performed by another examiner. This level of blinding would be inappropriate in a study focused on the EDX consultation of a person with low back pain. In that instance, the consultant might perform all of the testing, history, and physical examination, and only be blinded to a reference standard (e.g., the MRI findings).

It is also probable that the extent of blinding may vary with different disease categories and severity. For example, in a study of persons with severe congenital myopathy compared to asymptomatic volunteers, the physical appearance of the subjects unavoidably would introduce a bias. On the other hand, a study of low back pain volunteers compared to persons with mild radicular symptoms might be adequately blinded by simply not allowing the investigator to discuss symptoms with the subjects. This is a researchable question, and more work may define the extent of the blinding needed in certain circumstances.
Three possible options to eliminate bias in this area are remote, parallel, and tandem designs. (Figure 1) With the remote design, the researcher interprets the waveforms obtained in a standardized fashion outside of the clinical environment, without knowledge of clinical information. In the parallel design, the researcher performs the EDX study independent of the clinician and with no knowledge of the clinical history, and a diagnosis is made. The protocol is developed to minimize the transfer of clinical information to the investigator. In the tandem design, the researcher is present in the room and can direct the examination, but cannot see the subject.

CONCLUSION

To minimize bias and to improve the accuracy and objectivity of EDX techniques, future studies evaluating a diagnosis using EDX testing must meet two criteria. First, the results of the research test and the reference standard should not be known by the investigator (e.g., the EDX consultant interpreting the needle electromyography (EMG) study should be blind to the results of the history and physical examination). Second, the patients in the study should undergo both the research test in question (e.g., an EDX study to evaluate CTS); and the reference standard (the history and physical examination). In this way, investigators avoid the conscious and unconscious bias that might otherwise cause the research test to be interpreted as abnormal when the reference standard test is positive and interpreted as normal when the reference standard test is negative. Those involved in EDX research must adopt the current research standards of conducting blinded studies. Only through research meeting these standards will EDX physicians be able to establish the validity of EDX testing. If EDX physicians fail to adopt these concepts and to improve research techniques, they deserve to be left behind.

REFERENCES


The Use of Functional Outcome Measurements In Clinical Neuromuscular Disease Trials

David S. Saperstein, MD
Assistant Professor of Neurology
University of Kansas Medical Center
Kansas City, Kansas

INTRODUCTION

One of the most important and often most difficult aspects of designing a clinical trial is selecting the primary endpoint. The question of interest is usually whether or not patients have improved. How to best assess improvement depends on the disease being studied. For trials involving motor neuron disease or muscular dystrophy, muscle strength would be a reasonable outcome measure. In comparison, in a study of peripheral neuropathy, strength may be a meaningful outcome, but endpoints such as sensation might be appropriate. This is further complicated due to the various ways to assess improvements in sensation—such as decreased numbness, improved balance, or decreased pain. Similar difficulties are illustrated by imagining a clinical trial involving patients with myasthenia gravis (MG): improvement could be measured in terms of disparate functions such as changes in strength, fatigue, vision, swallowing, or breathing.

Regardless of the clinical function of interest, a reliable means for measuring change in this function is essential. This means of measurement must be amenable to assessment with statistical methods. In the routine care of patients, clinicians quantitate neurologic function, often employing a number of commonly used scales for strength, sensation, and reflexes. Such scales may or may not be suitable for a clinical trial. Choosing a measurement scale as an endpoint for a clinical trial requires consideration of several factors.

The ideal outcome measure should consider:

1. Whether it is **appropriate to the function** it is supposed to measure (i.e. does a test of motor function favor patients with greater preserved strength or will changes in weak patients be detected as well).
2. Whether the **measure is valid** (how does it compare with other measures designed to test the same function).
3. Whether it is a **reliable and reproducible** test (can the measure be repeated accurately, can it be reproduced by others, and is it simple enough to be repeated without significant additional training).
4. Whether it is **sensitive** to detect changes in the conditions under study, considering the variability arising from other aspects of the patient’s disability.

Often the best way to meet these requirements is by using a functional rating scale. These scales produce a score that incorporates function or performance in various spheres. Some scales may focus on measures of strength or sensation whereas others may focus on what activities a patient can perform (e.g., how far a patient can walk unassisted). The latter type of scale has the advantage of simultaneously considering multiple variables. Another role served by such a scale is that it typically provides
more useful information than just knowing that there has been a change in an examination finding, such as sensation or strength. Outcomes measured in terms of functional ability tend to be more meaningful to patients and healthcare providers. In clinical research, a measurement scale is often “borrowed” from one disease for use in another (such as using a functional scale from a stroke trial to study peripheral neuropathy). Validation of a measurement tool for a specific disorder is an important, but often neglected task.

Of primary importance regarding any measurement is minimizing error variance (the variability within a measurement). For example, if a true difference existed between a control and treatment group, but the measurement variable involved significant random error variance, the difference between groups could be missed. Error variance increases with the number of treatment groups and with the complexity of the measurement. Therefore, it is preferable to minimize the number and complexity of variables while maximizing the number of observations or patients.

Another important aspect of study design involves sensitivity. Obviously, an investigator would like to be able to detect as small as possible a difference between treatment groups. The main limiting factor is the number of subjects in the study. Cost and availability of subjects are the two main constraints on sample size (limited availability of subjects is a common problem in the study of rare conditions such as neuromuscular disorders). The number of subjects needed to be able to identify a reasonable degree of difference will vary depending on the primary outcome measure used. Therefore, the selection of a specific outcome measure is often governed by the number of subjects an investigator will be able to find or afford to recruit.

An essential distinction among measurement scales is whether the data measured are continuous or noncontinuous. Continuous measures have uniform intervals between scores. A 50% increment between scores at the low end of the measurement range implies the same relative change as a 50% increment at the high end of the measurement scale. The use of continuous measures usually implies that the values will be normally distributed (bell shaped or Gaussian) around a mean value. Examples of continuous data are weight, nerve conduction velocity, and force generated by muscle. Continuous data can be analyzed using parametric statistics.

Noncontinuous measures may be categorical or ordinal. Examples of categorical variables are gender, “improved,” or “weaker.” Ordinal data involve a rank order, with one score being better or worse than another. An ordinal scale does not permit the user to determine how far apart the points on a scale are, but only to hierarchically rank each point. Ordinal data are collected as discrete values. A common ordinal scale is the use of “severe, moderate, or mild” as a means of classifying the severity of disease. The 0 to 5 medical research council (MRC) scale for grading motor strength is another commonly used ordinal scale. When using ordinal data, the different scores cannot be compared in an absolute way (i.e., a difference between a score of 2 and 3 is not the same as a difference between a score of 4 and 5).

Ordinal scales must be analyzed using nonparametric statistics. These statistical comparisons tend to be less powerful than parametric statistics since only the relative relationship of one score to the other is taken into account. The magnitude of the differences between scores or the variability of scores are not considered. Parametric statistics tend to be better at detecting differences compared with nonparametric analyses. Nonparametric measures are, however, often easier to gather and are much less prone to error variance.

**MOTOR SCALES**

**Manual Muscle Testing**

Manual muscle testing (MMT) grades muscle strength on a scale of 0 to 5. It requires no equipment, all electrodiagnostic consultants should be familiar with it, and it can be performed quickly. In addition, good inter-evaluator reliability can be obtained. There are concerns that MMT might not be as sensitive to small changes in strength as more quantitative methods. Indeed, a decrease in as much as 40% of a muscle’s capacity for contraction could be graded with the same MMT score. However, a recent study suggests that if enough muscles are assessed, MMT can have the same or better sensitivity for decline in strength as quantitative methods (as discussed later). A distinct disadvantage of MMT is that it produces ordinal data. Since the intervals between ranks are not equal, statistical analysis must use nonparametric techniques that are less powerful and sensitive than the parametric statistics that can be used with continuous data.

**Quantitative Strength Testing**

**Maximal Voluntary Isometric Contraction**

Maximal voluntary isometric contraction (MVIC) uses computerized dynamometry to obtain objective measurements of force generated by different muscle groups. Force is measured in Newtons, thus data is continuous, avoiding the limitations of MMT with respect to statistical analysis. It had traditionally been held that MVIC is more sensitive to early changes in muscle strength than MMT, but this may not be the case. In clinical trials, MVIC scores are usually standardized as z-scores. Z-scores represent the ratio of the observed force minus the predicted force, to the standard deviation of the predicted force (the predicted force and standard deviation are calculated from a cohort of amyotrophic lateral sclerosis [ALS] patients). Each muscle has different normal force and changes in strength vary.
Z-score normalizes this change and allows averaging of different muscle groups into a “megascore.” Alternatively, MVIC data can be normalized to percentages based on gender, age, and height. Using MVIC excellent intra- and inter-rater reliability can be achieved. MVIC has been used as the primary outcome measure in a number of neuromuscular trials, especially treatment trials in ALS. There are conflicting data, but it appears that changes in MVIC scores are correlated with and can predict decline and survival in patients with ALS.

Disadvantages of MVIC are that testing is time consuming (approximately 45 minutes) and requires specialized equipment and trained personnel. Subjects must be accurately positioned in the testing apparatus. As patients become weaker, as in ALS trials, positioning and cooperation become more difficult.

A potential alternative to MVIC is hand-held dynamometry. This has been used in several neuromuscular trials. In ALS patients hand-held dynamometry has been shown to have a similar rate of change and shows only slightly more variability than MVIC. There is a high correlation with MVIC scores and similar test-retest variability. Like MMT, hand-held dynamometry can be accomplished in just a few minutes.

**PULMONARY SCALES**

Pulmonary function tests, especially forced vital capacity (FVC) are similar to MVIC in that they produce continuous variables and scores can be normalized to patient parameters, such as age and body size. With respect to neuromuscular disorders, FVC has been used primarily in ALS clinical studies. It is relatively easy to perform with minimal and portable equipment. A decline in FVC correlates with survival in ALS patients. However, measurements may be difficult to accurately obtain if there is facial weakness since patients may not be able to get a good lip seal around the mouthpiece. In addition, patient effort and cooperation can have significant effects on measurements. Forced vital capacity may be affected by factors not related to the underlying neuromuscular disease, such as a primary pulmonary process like asthma or chronic obstructive pulmonary disease. Another limitation of FVC is that it is not sensitive to early or mild weakness.

**DISABILITY SCALES**

The results of a clinical trial are often difficult to translate into a meaningful perspective. Suppose a study of drug X shows that the decline in strength of ALS subjects is decreased by 10%. It is difficult for clinicians and patients to appreciate the practical significance of these results. A more useful metric would be to show how a patient’s functional status is affected. To this end, functional or disability scales can be useful.

Measurements of disability are usually more meaningful to people’s lives than objective measures such as grip strength or timed walking. Unfortunately, a patient’s performance on a disability scale may depend on a number of factors. People with similar deficits may manifest different levels of disability based on motivation and support systems available to them. Another disadvantage is that disability scales produce ordinal data that must be handled with nonparametric statistical methods.

Disability scales have been used in a number of trials of ALS, peripheral neuropathy, and muscular dystrophy. A current trend in ALS research is to use disability scales as the primary outcome measure.

The Amyotrophic Lateral Sclerosis Functional Rating Scale

The Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) was designed to assess the ability of ALS patients to perform activities of daily living in order to determine if there is any functional improvement or decline during a study period. This is an easily administered ordinal rating scale measuring patients’ assessments of their ability and independence in 10 functional activities. The ALSFRS assesses four bulbar-respiratory functions, two upper extremity functions (cutting food and dressing), two lower-extremity functions (walking and climbing), and two other functions (dressing/hygiene and turning in bed) (Table 1). For each function, patients select from a list of five choices, scored from 0-4. The score ranges from 40 (normal function) to 0 (unable to attempt the task). The ALSFRS takes approximately 10 minutes to administer and this can be done by a trained assistant, therefore not requiring the time of a physician or nurse. Excellent intra- and inter-rater reliability can be achieved. The ALSFRS can even be reliably performed over the telephone.

The ALSFRS has been used extensively in ALS clinical trials and correlates well with quantitative strength testing results and functional measurements. A disadvantage is that the ALSFRS is insensitive to small differences.

**Rankin Scale**

The Rankin Scale was initially devised as a stroke research equivalent to the Glasgow Outcome Scale. The Rankin scale is an
Table 1  The Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS):

1. SPEECH
   4 Normal speech processes.
   3 Detectable speech disturbance.
   2 Intelligible with repeating.
   1 Speech combined with non-vocal communication.
   0 Loss of useful speech.

2. SALIVATION
   4 Normal.
   3 Slight but definite excess of saliva in mouth; may have nighttime drooling.
   2 Moderately excessive saliva; may have minimal drooling.
   1 Marked excess of saliva with some drooling.
   0 Marked drooling; requires constant tissue or handkerchief.

3. SWALLOWING
   4 Normal eating habits.
   3 Early eating problems, occasional choking.
   2 Dietary consistency changes.
   1 Needs supplemental tube feeding.
   0 Nothing by mouth (NPO) (exclusively parenteral or enteral feeding).

4. HANDWRITING
   4 Normal.
   3 Slow or sloppy; all words are legible.
   2 Not all words are legible.
   1 Able to grip pen but unable to write.
   0 Unable to grip pen.

5a. CUTTING FOOD AND HANDLING UTENSILS (patients without gastrostomy)
   4 Normal.
   3 Somewhat slow and clumsy, but no help needed.
   2 Can cut most foods, although clumsy and slow; some help needed.
   1 Food must be cut by someone, but can still feed slowly.
   0 Needs to be fed.

5b. CUTTING FOOD AND HANDLING UTENSILS (alternate scale for patients with gastrostomy)
   4 Normal.
   3 Clumsy but able to perform all manipulations independently.
   2 Some help needed with closures and fasteners.
   1 Provides minimal assistance to caregiver.
   0 Unable to perform any aspect of task.

6. DRESSING & HYGIENE
   4 Normal function.
   3 Independent and complete self-care with effort or decreased efficiency.
   2 Intermittent assistance or substitute methods.
   1 Needs attendant for self-care.
   0 Total dependence.

7. TURNING IN BED AND ADJUSTING BED CLOTHES
   4 Normal.
   3 Somewhat slow and clumsy, but no help needed.
   2 Can turn alone or adjust sheets, but with great difficulty.
   1 Can initiate, but not turn or adjust sheets alone.
   0 Helpless.

8. WALKING
   4 Normal.
   3 Early ambulation difficulties.
   2 Walks with assistance.
   1 Nonambulatory functional movement only.
   0 No purposeful leg movement.

9. CLIMBING STAIRS
   4 Normal.
   3 Slow.
   2 Mild unsteadiness or fatigue.
   1 Needs assistance.
   0 Cannot do.

10. DYSPNEA
    4 None.
    3 Occurs when walking.
    2 Occurs with one or more of the following: eating, bathing, dressing (active daily living).
    1 Occurs at rest, difficulty breathing when either sitting or lying.
    0 Significant difficulty, considering using mechanical respiratory support.
ordinal scale of functional outcome that uses 6 grades, from 0 (no symptoms at all) to 5 (severe disability; bedridden, incontinent, and requiring constant nursing care and attention). In most studies, patients are divided into two groups (0-3, 4-5) or three groups (0, 1-3, 4-5) in an attempt to quantitate outcome.

The Rankin scale has proven to be reliable and is simple to administer, making it well-suited for use in large multi-center clinical trials.\(^6\) Although it was designed to measure outcome in stroke patients, this scale has been employed in several neuromuscular trials.\(^5,14\) The major limitation of the Rankin scale is its inherent insensitivity. Objective and subjective assessments are combined as are as measurements of impairment and disability. The Rankin scale is heavily weighted towards mobility.

**Hughes Functional Disability Scale**

The Hughes Functional Disability Scale (HFDS) grades physical function on a 7 point scale: 0 = normal; 1 = minor symptoms capable of running; 2 = able to walk 30 feet without assistance but unable to run; 3 = able to walk 30 feet with the assistance of one person, a walker, or a cane; 4 = unable to walk; 5 = requires assisted ventilation; 6 = death.\(^11,20\) It is a simple scale that provides meaningful information with respect to physical function. Like other simple scales, its main limitation is an inability to detect small differences.

**Myasthenic Muscular Score**

The Myasthenic Muscular Score (MMS) consists of nine assessments ranging from timed tasks (e.g., maintaining arm extension) to functional activities (e.g., chewing). For the timed tasks, points are assigned based upon how long a task can be performed. The functional activities are graded using an arbitrary ordinal scale. The MMS has been used in two multicenter randomized trials.\(^14,22\) This scale can be administered in 5-10 minutes and requires no specialized training.

The reproducibility of the MMS has not been evaluated. In the two trials in which this scale was used as a primary endpoint, no difference was found between the treatment groups, suggesting this instrument might not be sensitive to small changes in clinical outcome.

**Quantitative Myasthenia Gravis Score**

The Quantitative Myasthenia Gravis score (QMG) was initially developed by Besinger and Toyka in 1981 and later modified by Tindall\(^7\) and Barohn.\(^8\) The QMG consists of several independent assessments including: (1) time to develop diplopia with prolonged lateral gaze; (2) time to develop ptosis with prolonged upgaze; (3) facial strength; (4) ability to swallow 1/2 cup of water; (5) speech following counting aloud from 1-50; (6) time patient can outstretch either arm; (7) vital capacity; (8) right and left grip strength; (9) time patient can lift head 45 degrees in a supine position; and (10) time patient can outstretch either leg.

In contrast to most of the scales discussed, good inter-examiner reliability has been proven. The QMG has been used in a therapeutic trial of cyclosporine in MG,\(^31\) and is a major outcome measure in ongoing trials of mycophenolate mofetil in MG.

However, there are a number of disadvantages to the QMG. A patient can have an improved total score, but be functionally limited by weakness in one or two areas. Therefore, the QMG does not necessarily reflect the patient’s functional status. In addition, completing this scale is time consuming.

**Myasthenia Activities of Daily Living Score**

The Myasthenia Activities of Daily Living Score (MG ADL) score is an example of a disease-specific functional status measurement score. It consists of eight functional tasks which are scored on a scale from 0 (normal) to 3 (severe impairment): (1) talking; (2) chewing; (3) swallowing; (4) breathing; (5) impairment of ability to brush teeth or comb hair; (6) impairment of ability to arise from a chair or toilet; (7) double vision; and (8) eyelid droop. The MG ADL correlates well with the QMG.\(^33\)

The MG ADL has been adopted in several MG clinical trials. It is simple and economical to administer and does not require an extensive amount of time to complete. The main disadvantage is that the responses are largely subjective.

**CONCLUSIONS**

As illustrated by this manuscript, a number of clinical measurement scales have been employed in clinical trials of neuromuscular disorders. When selecting a scale as an endpoint in a clinical trial, consideration must be given to whether the scale measures an appropriate clinical outcome. In addition, this measurement must be valid, reliable, and sensitive to small changes. These factors must be balanced against the number of subjects needed to measure a meaningful change, the ease with which measurements can be made, and how amenable the data are to statistical analysis.

**REFERENCES**


Electrodiagnosis for Prognosis: Clinical Research Opportunities

James M. Gilchrist, MD
Professor of Neurology
Department of Clinical Neurosciences
Chief, Division of Neuromuscular Disease and Electromyography
Brown Medical School
Providence, Rhode Island

INTRODUCTION

The use of electrodiagnosis in the prognosis and management of neuromuscular disorders has recently been reviewed. In a discussion about properly developing and conducting clinical research trials of electrodiagnostic (EDX) procedures, this may not initially seem relevant. But the literature on prognosis of neuromuscular disease in general is inadequate, and in particular, electrodiagnosis is even more bereft of salient studies, especially those with the primary purpose of clarifying prognosis or management. Thus, the use of electrodiagnosis for prognosis and management would seem fertile ground for clinical research, as it has been left largely unplowed. The purpose of this manuscript will not be the how and why of clinical research design, but rather a proferring of potential research opportunities in an area usually neglected. This author submits that few if any of the studies that will be discussed have had rigorous research designs, and are therefore worthy candidates for study under the guidelines delineated by the other information that will be presented.

There are substantial limitations to the use of electrodiagnosis for prognosis, but only two of these limitations will be discussed. This author hastens to add that neither of these limitations are insurmountable barriers. The primary hurdle is the inherent limitation of EDX techniques to physiologic measurement. For the same reason that electrodiagnosis may be of limited utility if the pathology is not physiologic, “electroprognosis” will be limited if the outcome is not determined by physiologic factors. For example, EDX findings correlate poorly, if at all, with sensory symptoms, such as numbness and most especially, pain. This is evident when reviewing the literature on prognosis in radiculopathy, entrapment neuropathies, and diabetic neuropathy, in which pain, numbness, and other sensory manifestations largely determine outcome.

Despite the extensive literature on electrodiagnosis of radiculopathy, few articles address prognosis or longitudinal management. One needle electromyography (EMG) study that tracked spontaneous activity in paraspinal and limb muscles in cases of cervical radiculopathy failed to show any correlation between symptom duration and the occurrence of fibrillation potentials or positive waves. With this in mind, it is not surprising that other studies have found EMG of little help in predicting symptomatic outcome in radiculopathy. Falck and colleagues evaluated 80 cases of acute lumbar radiculopathy with EMG at onset and then at 1 and 5 years following hospitalization or surgery. Outcome was assessed by a modified World Health Organization handicap classification. In both surgically and conservatively managed patients, EMG had less prognostic value than other clinical and psychosocial factors. In a study of 37 patients treated surgically for lumbar radiculopathy, Rodet and colleagues found preoperative EMG did not correlate well with the patients’ subjective global assessment of outcome. Similarly, preoperative EMG failed to predict surgical outcome in a study of 26 patients with cervical radiculopathy. The prognostic utility of EMG is likely limited by its poor correlation with pain, a major determinant of disability and outcome in radiculopathy.
The role of EDX testing in prognosis and management of entrapment neuropathies is also complicated by poor correlation between the degree of EDX abnormality and the severity of symptoms. Pain and paresthesias, often the earliest and most distressing symptoms, cannot be assessed with routine nerve conduction studies (NCSs) or EMG. Just as symptomatic entrapments may appear normal with routine EDX studies, conduction slowing or mild denervation may be clinically silent. A number of studies have attempted to assess the value of pre-operative EDX testing in predicting response to carpal tunnel release.\(^2\) No clear consensus has emerged from these retrospective reviews. When EDX studies did correlate with prognosis, the finding of moderate prolongation of sensory and motor latencies, preserved compound muscle action potential (CMAP) amplitude, and lack of denervation tended to predict favorable surgical outcome.\(^3,13,14\) The lack of more consistent correlation between EDX data and outcome may reflect that many outcome rating scales depend heavily on pain assessment.

Longitudinal studies including asymptomatic and symptomatic diabetics have indicated that nerve conduction data alone are of limited prognostic value.\(^7,8,22\) On the other hand, NCSs provide a sensitive and reproducible means of following diabetic polyneuropathy and evaluating experimental treatments. They present certain limitations including insensitivity to small-fiber sensory and autonomic function. Furthermore, nerve conduction abnormalities may not correlate closely to symptom severity or neuropathic impairment in some instances. One potential way to side step this disconnect between symptoms and neurophysiology is the use of composite or cross-modality indices comprised of various nerve conduction parameters and, conceivably, other physiologic and non-physiologic measures. A recent report\(^8\) concerning diabetic neuropathy favors composite scores of multiple nerve conduction parameters as more reproducible and sensitive than individual attributes. For diabetic neuropathy, their results demonstrated that composite scores of velocities and latencies were particularly sensitive to the presence of neuropathy, whereas composite scores of amplitudes showed better correlation with measures of neurologic impairment.

A second limitation to the use of electrodiagnosis for prognosis is the relatively poor resolution of abnormalities which may not allow differentiation of outcomes. The lack of resolution in the description of EDX abnormalities is most acute when examining EMG parameters in the prognosis or management of myopathies, in large part because of the inability to correlate measures of motor unit potential morphology and recruitment, and spontaneous activity with severity of disease. The lack of specific systems of qualifying abnormalities noted on EMG, beyond noting their presence on a crude 1 to 4 scale, has hampered the use of EMG for prognosis, as fine grades of difference are lost and serial comparisons tend to be insensitive. Quantitative EMG (QEMG) may be a technique to circumvent this by allowing for a continuous, ordinal scale of measurement. However, at this point, QEMG studies have been ambiguous on the issue of progression over time.\(^12\) Quantitative EMG also provides the ability to accurately replicate measurement techniques and allow serial comparisons, as well as correlations to non-physiologic measures,\(^2\) and therefore holds promise in providing useful information for prognostication. Single-fiber EMG (SFEMG) is another quantitative technique, but one in which the relevant physiology is not particularly attuned to myopathies. Jitter reflects ongoing changes at the neuromuscular junction related to necrosis of myofibers and subsequent regeneration, both important features of some myopathies, but as yet not shown to provide prognostic information.

Despite the limitations discussed, there are disorders in which prognosis is the defining purpose of the EDX study, in large part because physiology plays an important role in determining outcome. Two examples are Bell’s palsy and traumatic nerve injury. In the case of Bell’s palsy, electrodiagnosis has been a partial success. Numerous EDX methods can be used to study the facial nerve, including facial CMAP amplitudes (also known as evoked EMG and electroneurography) and latencies, EMG of facial-innervated muscles, the blink reflex, antidromic facial nerve action potentials, minimal nerve excitability testing, maximal nerve excitability testing, and transcranial magnetic stimulation (TMS).\(^12\) At this time there is no EDX technique that reliably predicts prognosis in Bell’s palsy within the first 24-48 hours after onset, i.e., at a time that may influence treatment options. Compound muscle action potential amplitude comparing side-to-side difference at day 5 to 7 after onset appears to be the most reliable parameter for ultimate prognosis. When the CMAP amplitude is less than 10% of that on the healthy side, maximum recovery will be delayed 6-12 months and function will be moderately or severely limited. If the amplitude is 10-30% of the healthy side, recovery may take 2-8 months with mild to moderate residua. If the CMAP amplitude is greater than 30% of normal, complete recovery can be expected at 2 months after onset.\(^23\) Compound muscle action potential amplitude is most accurate in predicting a good prognosis, but less accurate in those patients with greater than 90% loss of amplitude, as a significant number (up to 47%)\(^30\) will still have a good recovery. The search remains for a neurophysiologic technique which can determine axonal loss prior to Wallerian degeneration.

Electrodiagnostic evaluation figures prominently in the approach to traumatic nerve injury. Initial studies, typically performed within the first month after injury, not only serve to confirm diagnosis but also provide important information for prognosis and management. One of the major roles of initial EDX studies is to help grade the severity of any nerve lesion. The classic scheme of Seddon\(^29\) grades nerve injuries as neurapraxic, axonotmetic, and neurotmetic is based on morphologic criteria, but also conveys important prognostic and therapeutic implications, in large part because the morphologic changes have par-
Serial SFEMG studies of patients with MG can also provide predictive value for worsening and for improvement. Sanders and colleagues\textsuperscript{27} found a tighter correlation between improvement and jitter reduction than between worsening and jitter increases, but nevertheless concluded that all three jitter parameters (individual fiber pair jitter, mean jitter, and percentage of blocking fibers) reflected disease change, for better or for worse. In two-thirds of patients with clinical worsening between SFEMG studies, there was at least a 10% rise in mean jitter. The 83% of patients who improved had at least a 10% decrease in mean jitter, and 81% of patients whose mean jitter fell by at least 10% had clinical improvement, a percentage which increased to 86% when there was a 20% or greater drop in mean jitter. The authors concluded that a decline in jitter of 10% or more had a strong correlation with clinical improvement.\textsuperscript{27}

Situations in which the potential exists for expansion of the role of electrodagnosis in prognosis and management include amyotrophic lateral sclerosis (ALS), traumatic nerve injury, root avulsions, obstetric brachial plexopathies, chronic inflammatory demyelinating polyneuropathy (CIDP) and polymyositis.

A number of studies underscore the value of motor unit number estimation (MUNE) in following ALS. Serial MUNE evaluations appear to be a more sensitive measure of disease progression, showing a steeper rate of decline than quantitative strength testing, forced vital capacity, or the Appel ALS rating scale.\textsuperscript{13} In a longitudinal assessment of hand muscles in ALS, MUNE (along with fiber density) was a more sensitive indicator of progression than CMAP amplitude or grip strength.\textsuperscript{40} Furthermore, rapid decline in MUNE appeared to indicate a worse prognosis whereas increasing fiber density suggested a more favorable course. A more recent study emphasized the prognostic value of MUNE, demonstrating that it predicted survival more accurately than serial assessment of strength or forced vital capacity.\textsuperscript{14} Given the paucity of established therapeutic options, MUNE currently is likely to be of greater use in clinical trials than in standard management of ALS patients. Ongoing clinical trials include MUNE as a secondary outcome measure. As these trials progress, the practical advantages or shortcomings of MUNE should become more clearly defined.

Transcranial magnetic stimulation has been investigated as a means of assessing upper motor neuron pathology in ALS but, to date, no studies have specifically evaluated the prognostic role of TMS in ALS or the ability to differentiate cases of primary lateral sclerosis from ALS with predominant upper motor neuron involvement (a distinction that would have clear prognostic implications).\textsuperscript{37}

Several aspects of traumatic nerve injury should be amenable to further neurophysiologic investigation of prognosis. It has been proposed that a small percentage of surviving motor units can restore considerable strength to a severely denervated muscle.
through collateral and terminal sprouting, though this has not been explicitly documented in longitudinal studies.\textsuperscript{29} Studies disproving this contention might support more aggressive management of nerve injury despite evidence of nerve continuity.

As mentioned earlier, differentiation between neurotmesis and axonotmesis is a critical determinant between conservative and surgical repair. Current methods often require 3 to 4 months before this can be ascertained, and even then it is often an evasive determination. Surgical exploration with intraoperative NCSs can be quite helpful but are obviously invasive. Development of techniques to make this differentiation earlier remains a holy grail of sorts for neurophysiologists. Ultimately, the most desirable test would be one which allowed determination of morphologic continuity at 1 to 2 days, when not only traumatic nerve injury but also Bell's palsy would be most amenable to surgical and other treatment interventions.

Lastly, aberrant regeneration is another source of poor outcome in traumatic nerve injury and Bell's palsy, leading to functional and cosmetic disability. This is particularly frustrating when nerve regeneration is otherwise quite robust. The detection of aberrant regeneration in the limbs relies heavily on EDX testing. The reinnervation of hand muscles is particularly susceptible to aberrant regeneration as demonstrated by Thomas and colleagues.\textsuperscript{33} They performed EDX assessments of 11 patients following repair of ulnar and median nerve transactions at the wrist as well as in 2 patients with repaired median nerve transactions at the elbow. Careful needle EMG revealed motor units in intrinsic hand muscles recruited primarily during voluntary activation of other hand muscles. In addition, motor unit recruitment did not follow the normal size-ordered sequence, as demonstrated through spike-triggered averaging. These abnormalities, attributed to misdirected motor units, likely contributed to the poor recovery of hand dexterity in many of the patients. In adults, compensation for aberrant regeneration appears limited so EDX evidence of misdirected motor units adversely affects the long-term prognosis.\textsuperscript{32} Electrodiagnostic investigations indicative of a high risk of aberrant regeneration early in the course are lacking but would be useful in management of traumatic nerve injury and Bell's palsy.

The detection of cervical root avulsion is of major importance as it precludes spontaneous regeneration and surgical re-anastomosis. Electromyography of paraspinal muscles lacks sensitivity due to overlap of segmental innervation and the possibility that paraspinal denervation may represent less severe root damage with potential for recovery. Preserved sensory nerve action potentials within anesthetic dermatomes or completely denervated myotomes is considered the strongest evidence of root avulsion. Trojaborg\textsuperscript{36} compared EDX data (including extensive sensory conduction studies) to results of myelography in 17 patients with presumed cervical root avulsion. The EDX and imaging studies concurred in identifying root avulsions for 8 of the cases. Among the other 9 cases, 5 showed EDX but not radiographic evidence of avulsion, 2 showed radiographic avulsion not confirmed by electrodagnosis, and 2 had avulsion localized to different roots by the 2 methods. Follow-up studies identified false positives for both methods. Somatosensory evoked potentials can identify traumatic root avulsions but may overdiagnose root avulsion.\textsuperscript{12} Further investigation is warranted.

Electromyography performed in the first months after a birth injury to the brachial plexus often provides an overly optimistic picture\textsuperscript{36,39} and there is a need for further refinement of electrodiagnosis for prognostic purposes. Muscles with many recruitable motor units frequently remain weak or paralyzed. A number of explanations have been proposed. First, the actual strength of muscles may be underestimated because of difficulty in examining infants or guarding from discomfort.\textsuperscript{36} Furthermore, an apraxia or proprioceptive deficit may supervene.\textsuperscript{5,28} Another explanation involves the geometry of myofibrils in immature muscles. The small, tightly packed fibers could produce a deceptively dense EMG interference pattern.\textsuperscript{35} Finally, muscles may contain aberrant motor units that contribute to EMG activity but not to effective or coordinated movement.

Because patterns of progression, recovery, and relapse are often complex in CIDP, prognostication is considerably more difficult than in AIDP. The degree of axonopathy remains a major determinant of prognosis. Distally evoked CMAP amplitudes provide an approximate assessment of axonal degeneration but will underestimate severity in longstanding cases involving terminal sprouting. The pattern of demyelination seen on nerve conduction studies has been associated with prognosis in CIDP on by Kuwabara and colleagues\textsuperscript{19} based on the presence of distal, intermediate, or diffuse demyelination. Distal demyelination was defined as distal motor latencies exceeding 125% of the upper normal limit. Intermediate demyelination was defined as motor conduction velocity less than 80% of the lower normal limit or a drop in CMAP amplitude greater than 20% over the forearm or calf. The distal pattern correlated with a monophasic, remitting course generally responsive to treatment. A more chronic, insidious course with less response to treatment characterized the intermediate group. Diffuse demyelination tended to occur in patients with long, relapsing courses. These results need to be confirmed.

Electromyography plays a major role in the diagnosis of inflammatory myopathies and potentially could provide a degree of prognostic assistance but has been underutilized for this purpose. In part, this is due to obstacles previously delineated such as non-neurophysiologic determinants of outcome and inadequate clarification of EMG abnormalities. But EMG abnormalities, particularly motor unit potential changes and fibrillations, correlate with abnormalities on muscle biopsy in polymyositis with
worse EMG findings indicative of more severe muscle pathology. Electromyography offers a window into the ongoing inflammatory process which is less invasive than a muscle biopsy and could potentially provide prognostic and management information of value to the clinician. The presence or absence of abnormal spontaneous activity can be a useful measure of response to treatment, with those patients with persistent fibrillation potentials presumably responding less satisfactorily than those patients in whom these potentials disappear. The occurrence of fibrillation potentials can also be used to assist in the management of patients with polymyositis who have been treated with steroids and are now weakening. The diagnostic dilemma is between unresponsive or recurrent inflammatory myositis and steroid-induced myopathy, a non-destructive process not associated with abnormal spontaneous activity. An EMG showing fibrillation potentials in weak muscles would support inadequate treatment of the myositis, whereas an EMG without such activity would support the need for further reduction in steroid dose. As stated above, QEMG may also provide a way to reliably predict outcome, overcoming the insensitivity of routine needle EMG.

**SUMMARY**

In summary, the capacity of electrodiagnosis to provide prognostic information on a particular disease is primarily dependent on its ability to measure the salient features determining outcome. The underlying factors responsible for outcome must first be defined, whether they be pathologic, immunologic, sociologic, psychologic, pharmacologic, or physiologic. In those diseases in which disordered physiology determines outcome, EDX may then be a good prognostic indicator, but only if that disordered physiology can be measured using EDX. In AIDP and Bell’s palsy, the outcome is not determined by the most prevalent pathologic feature, i.e., demyelination, but rather by axonal loss: those without any will do well, those with much will not. Thus, CMAP amplitude, a good measure of axonal physiology, becomes an accurate measure of prognosis. On the other hand, in polymyositis, the primary pathophysiology causing weakness is myonecrosis but the real indicator of outcome is the response of the inflammatory process to immuno-modulation, which cannot be measured by EDX techniques. Even indicators of myonecrosis, such as fibrillations, may not accurately portray prognosis. The mediocre value of NCS in predicting outcome from carpal tunnel syndrome surgery is largely because outcome is based not on physiology but on subjective aspects such as pain and numbness, which do not correlate well with EDX findings. Single-fiber EMG remains an imperfect prognostic tool in MG because outcome is largely determined by generalization of disease and response to immuno-modulatory therapy. Neither is well-portrayed by measurement of SFEMG jitter, as illustrated by the failure of SFEMG to accurately predict which patients will progress from ocular to generalized disease, and the lack of evidence that SFEMG jitter predicts which patients will respond to treatment.

**References**

Development and Use of Quality-of-Life Tools

Luca Padua, MD, PhD
Istituto di Neurologia
Università Cattolica S.C.
Rome, Italy

Irene Aprile, MD
Istituto Pro Iuventute Don C. Gnocchi
Rome, Italy

BACKGROUND

Traditional assessment of the outcomes of neurological and musculoskeletal conditions has been based on physician-derived and instrumental findings and in most cases these parameters are standardized. Furthermore, when physicians do look at the patient’s point of view, they do not usually do it in a standardized way and they do not rigorously "measure" the patient’s perceptions.

Over the last 2 decades, clinical and public health researchers have outlined the need for a thorough evaluation of concepts such as health-related quality of life (QOL) and patient satisfaction that strongly correlate with the patient-expected outcome. The need for standardized measures of such items has stimulated an extensive and rigorous process which has led to the development of validated patient-oriented instruments. These tools, mainly self-administered questionnaires, concentrate on how illness affects the individual and they permit physicians to quantify these new parameters with measures that are not influenced by the physician’s opinion. A worsening of QOL score usually arises when there is a discrepancy between physical ability and need; it is dependent on whether there is an actual or perceived need to perform a specific function and on the patient’s expectations and motivation.

If the goal is to help patients, physicians must know what patients want and how much they value what they want. Physicians must also be sure that those things can be measured with the reliability, sensitivity, validity, and responsiveness associated with the traditional objective measures they have been comfortable with until now. Several self-administered questionnaires quantitatively measure function and QOL and fulfill the measurement properties for evaluation research. The patients’ perspective is therefore translatable into numbers available for statistical analysis. It is fundamental that the questions are crafted carefully and tested for performance in a target population prior to clinical use, just as any machine, camera, or other device should be tested prior to general release. This is what clinical epidemiology deals with—providing clinicians with the instruments to evaluate the patient’s point of view in a scientific matter. In many medical and surgical areas (oncology, cardiology, etc.), outcome assessment for various conditions has already been performed and investigated using patient-oriented tools and is widely accepted.

Neurological and musculoskeletal diseases usually involve the general health status of patients, therefore the QOL may be strongly affected. Moreover, often these diseases are chronic and QOL issues are more important in such conditions. For this reason, the patient may provide more significant information with these disorders than in other areas of medicine. Recently these measures have been included in some neurological and musculoskeletal studies and have provided interesting results. The patient-oriented findings in these fields are useful to assess the natural history of many chronic diseases, to disclose significant differences between similar pathologies (particularly the impact of the disease on the patient’s life), and to provide data for improving therapeutical strategies in clinical
conditions that often may be treated with strong therapies, including their possible side effects.\textsuperscript{22}

The recently advocated standardization of diagnostic and therapeutic approaches should be based on the following: to compare different studies in the same field, to compare studies in different fields, to have common measures, to perform multicentric research, and to perform multi-perspective assessment. Patient-oriented tools may be important because one of their features is that the measures have the same value in different situations and in different countries (although in the latter case a rigorous process of translation/validation must be performed).

In order to obtain a suitable multi-perspective assessment of the diseases, the use of validated and standardized patient-oriented measurements should be associated to further standardization of the traditional outcome parameters.

Two words of caution are necessary. First, the patient-oriented evaluation must be an additional standardized perspective for evaluation of the outcome and should not reduce the importance of other traditional parameters. This set of research methodologies really represents an extension of the kind of clinical research that is familiar to all physicians,\textsuperscript{16} and it must be considered an evolution of the traditional multidimensional evaluation of clinical disorders and in no way reduce the physician-oriented and instrumental findings. Second, certain kinds of pathologies that involve cognitive function cannot be evaluated by these instruments as they stand now. This author believes that a valid patient-oriented approach associated with standardized traditional parameters provides a useful multiperspective assessment that should be performed in neurological and musculoskeletal research whenever possible.

**SPREAD OF AND INTEREST IN QUALITY-OF-LIFE ASSESSMENT AMONG NEUROLOGISTS**

Although many researchers have outlined the need of QOL assessment, and although an exponential increase of interest on QOL is observed (demonstrated by the exponential increase of articles including QOL assessment), few studies have assessed the perspective on QOL from a sample of physicians.\textsuperscript{3} In June 2002, the Italian Society of Neurology (SIN) for the promotion of the spreading of knowledge in the field of QOL assessment, founded the Italian Quality of Life Study Group.\textsuperscript{13} The group created a fact-finding study evaluating the spread and the degree of knowledge in the QOL field among neurologists.

The study was conducted using a questionnaire distributed during the XXXIII national meeting of the SIN. The questionnaire was administered to a random sample of the SIN congress participants and 350 subjects completed the questionnaire.

To establish whether the enrolled sample was representative of the population of congress participants, the average age of the enrolled sample and the population of the participants who gave their data for Continuing Medical Education credits (subjects 1406: mean age 44.1, standard deviation 8.3, range 25-70) were compared and a significant difference between the two groups was not shown. Results are summarized in Table 1.

The majority of the enrolled sample (91\%) indicated that it would be important either to increase the knowledge of the real impact of pathology on the patient’s QOL, or to better evaluate the therapy effect. The age of the attendee was not a factor on any of the items on the questionnaire.

That study established that in the examined sample of Italian neurologists there is a good spreading of knowledge in the QOL field. Moreover, researchers are aware of the importance of QOL and of the need to increase the knowledge of it in order to improve clinical practice and research.

**DEVELOPMENT OF DISEASE-SPECIFIC QUALITY-OF-LIFE TOOLS**

It is well known that disease-specific tools are more sensitive than generic measures.\textsuperscript{7} Previous studies\textsuperscript{18,20} evaluating QOL in patients with myasthenia gravis (MG) through application of the most used generic measure, Medical Outcomes Study Short Form (SF-36), have demonstrated that QOL assessment is important in patients with MG. However, the Task Force of the Medical Scientific Advisory Board of the Myasthenia Gravis Foundation of America (MGFA) suggested developing a disease-specific questionnaire for the assessment of outcome measure in this pathology.\textsuperscript{9}

This suggestion led to the development and validation of a disease-specific questionnaire for MG called the Myasthenia Gravis Questionnaire (MGQ).\textsuperscript{14} The development and validation of a QOL instrument is standardized. This author’s approach was consistent with previously described strategies for scale development. There were three stages to scale development: Stage I - item generation; Stage II - item reduction; and Stage III - reliability and validity (Table 1).\textsuperscript{5}

In Stage I, item generation, a group of methodologists and clinical experts generated a list of 56 items. The list was based on (1) a previous study on MG sampling,\textsuperscript{18} (2) clinical experience, and (3) items proposed by MG patients. In Stage II, reduction of items, this 56-item list was reduced on the basis of results from field testing (41 patients completed the 56-item questionnaire). In Stage III, reliability and validity were assessed. The questionnaire provided a measure of the functional status and not of the mental picture (disease-specific measurement is not...
necessary for mental assessment). Results from field testing (prospective multidimensional study on 41 MG patients) showed that the final version of the questionnaire (25-item scale) was closely related to most findings at clinical examination. Single items concerning daily activity that need global motor function and the “global MG score” appeared strongly related to the muscle strength of the proximal muscles. Even items concerning work activity appeared related to clinical findings for proximal muscles. Clinically assessed bulbar innervated muscle function was strongly related to many items, particularly those connected with speaking or singing. Ocular symptoms appeared to be correlated to only two items concerning the activity of seeing. In a further phase, if subscore-domains are developed, one will be able to evaluate if the items are sufficient to measure the impairment in patients with ocular MG, otherwise it will be necessary to increase the number of items concerning these ocular symptoms. All but one of the single items (seeing limitation due to diplopia) and the global MG score were strongly related to the Osserman grade. Note that diplopia for the construction of the Osserman and MGFA classifications is not theoretically related to the scale (the first class includes pure ocular forms).

The MGQ had excellent internal consistency, with a Cronbach alpha of 0.95. Similarly, reproducibility of the questionnaire was good. In the present study, the test-retest interval was relatively short (a mean of 2 days) in order to limit the possibility of true clinical change that may falsely reduce the reproducibility (MG is a disease with marked frequent fluctuations). This may, however, increase the chance of memory effect, falsely improving the measured reproducibility of the questionnaire. For this reason the order of the questions was different in the retest version of the questionnaire.

In conclusion, the MG questionnaire provided a reliable and valid perspective measure of the functional status of the daily activity of myasthenic patients.

### DEVELOPMENT OF SUBSCALE-DOMAINS

After the first study, this author’s group evaluated the possibility of aggregation of the items into subscale-domains. In fact, in order to better assess generalized bulbar and ocular involvement, some items of the questionnaire were aggregated and validated into three domains as outcome measures. Construct validity was assessed by correlation of the domain results and conventional MG measurements; moreover, responsiveness was demonstrated. A wider and multicentric sample was assessed. Ninety-three MG patients (mean age 60.9, range 17–87, 41 males, 52 females) were prospectively enrolled in neurological departments of two neuromuscular transmission disease centers; furthermore, 23 MG patients were studied after initial treatment to evaluate responsiveness. Results from domains appeared related to conventional measures of MG severity. Questionnaires appeared more sensitive than clinical examination and history to assess bulbar and ocular involvement. The MGQ domains appeared reliable and sensitive.

In conclusion, the questionnaire, used in a large sample of patients, showed that the global MG score of domains appeared

<table>
<thead>
<tr>
<th>Stage</th>
<th>Methods</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Item generation</td>
<td>Regrouping of items generated by clinical experts, by MG patients, and some items inspired by SF-36.</td>
<td>List of 56 items</td>
</tr>
<tr>
<td>2: Item reduction</td>
<td>Administration of the 56-item questionnaire and clinical assessment in 41 MG patients. Analysis of item responses using statistical reduction techniques. Generation of MG score.</td>
<td>List of 25 items</td>
</tr>
<tr>
<td>3a: Reliability</td>
<td>Analysis of the correlation between MG score and conventional clinical measurements.</td>
<td>Significant correlation: reliability</td>
</tr>
<tr>
<td>3b: Validity</td>
<td>Tests of reproducibility, responsiveness, internal consistency, etc.</td>
<td>Reliability of all tests</td>
</tr>
</tbody>
</table>

A list of studies in which the author’s group correlated the neurophysiological findings with the patient-oriented results.

MG = myasthenia gravis; SF-36 = Medical Outcomes Study Short Form.
reliable. Domains were able to better assess the functional involvement of anatomical districts than the main global MG score, clinical examination, and history.

POLYNEUROPATHY AND QUALITY OF LIFE

This author's group performed a study\(^1\) that focused on the relationship between a patient's perception of his own inferior limbs' symptoms and function, and the clinical-neurophysiological assessment in patients affected by insulin-dependent diabetes mellitus (IDDM). The group studied 50 consecutive outpatients affected by IDDM by using validated measurements: (1) clinical (Semmes-Weinstein, vibration perception threshold, muscle strength, osteotendineous reflexes, etc.), (2) neurophysiological (sural, peroneal nerves), (3) metabolic, and (4) patient-oriented/QOL (North American Spine Society questionnaire). It was observed that patient-oriented scores were significantly related to neurophysiological findings of the inferior limb. Conduction velocity of the peroneal nerve was significantly related to symptoms and function of inferior limbs (p<0.05, r=0.3). In other words, patients with lower peroneal nerve conduction velocity more frequently reported disability and pain of inferior limbs. Sural nerve findings and the amplitude of peroneal motor response were not related to the patient-oriented parameters.

The data suggests that electrodiagnostic tests are useful to assess the severity of the diabetic polyneuropathy not only because they provide a biological measurement of the nerve function, but also because they appear related to the patient's QOL regarding peripheral nerve involvement.

SUMMARY

Traditional assessment of the outcomes of neurological and musculoskeletal conditions has been based on physician-derived and instrumental findings, but over the last 2 decades, clinical and public health researchers have outlined the need for a thorough evaluation of concepts such as health-related QOL. If the goal is to help patients, physicians must know what patients want and how much they value what they want. Several self-administered questionnaires quantitatively measure function and QOL. The patients' perspective can therefore be put into numbers available for statistical analysis.

Neurological and musculoskeletal diseases usually involve the general health status of patients, therefore the QOL may be strongly affected. For this reason, this kind of patient may provide more significant information. Recently these measures have been included in some neurological and musculoskeletal studies and have provided interesting results.

REFERENCES


