Experimental Strategies for Research on Multiple Chemical Sensitivity

Bernard Weiss

University of Rochester School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, New York

Skepticism about the validity of the multiple chemical sensitivity (MCS) syndrome stems in part from the lack of supporting experimental data. Performing the relevant experiments requires investigators to take account of broad variations in sensitivity and the need to establish reproducibility. The research approach best suited for MCS studies is the single-subject design. In contrast with conventional group designs, such designs emphasize repeated observations on individual subjects. Repeated observations of this kind constitute a time series in which successive measurements are serially or autocorrelated. One statistical method that bypasses the serial correlation problem is randomization tests. Explicit time series analyses take account of this aspect and can correct for it to determine the impact of an intervention such as a chemical exposure. — Environ Health Perspect 105(Suppl 2):487–494 (1997)

Key words: multiple chemical sensitivity, randomization tests, time series analysis, single-subject designs

The pivotal question posed to the members of the biomedical community is the authenticity of the multiple chemical sensitivity (MCS) syndrome. Do they accept it as a valid clinical entity? Most clinicians and biomedical scientists remain dubious, a point of view reflected in the American Medical Association's (AMA) position paper on clinical ecology (*1*). Their positions are not likely to be dislodged except by a compelling mass of cumulative evidence. If such evidence is offered, it must be a product of investigations that meet recognized standards of experimental and epidemiological design. Case reports and testimonials, no matter how numerous, usually fail to convince skeptics. Conventional case studies cannot confirm the MCS hypothesis.

One of the impediments to experimental verification of MCS and the source of much of the skepticism aroused in the biomedical community is the skepticism toward research shared by many clinical ecologists. Most crucially, they are not inclined to perform the kinds of experiments that scientific investigators find convincing. Further, they are wary of experiments conducted by others that do not completely accept their premises; they pose objections that make it arduous if not impossible to conduct clinical trials based on double-blind, placebo-controlled designs. For these disciples, the tenets of clinical ecology require that patients reside in what they call an environmental unit, a place with an environment devoid of volatile organics from furniture, walls, synthetic fabrics and other sources, and free of pesticides, drugs, and other factors that may induce reactions on exposure. Drinking water sources are also restricted. Fasting for several days in such a unit is also recommended or required until the patient's symptoms are deemed to have cleared. Only then, with the reactions unmasked, is it considered possible to test the response to an acute exposure.

Data are not readily available on how much chemical purity such units achieve. Advanced analytical methods would be necessary to confirm the absence of chemical, microbial, and physical agents implicated as potential stimuli. For example, concentrations of particulate matter, especially particle-size distributions, would need to be established if the claims of what is achieved by environmental units are granted credibility. Because it is seen as a possible source of adverse responses, drinking water in such units should meet or even exceed the U.S. Environmental Protection Agency's (U.S. EPA) required sensitivity standards. In contrast to the extreme positions described above, other clinical ecologists believe that empirical outpatient treatment can achieve acceptable clinical results, bringing patients to the unmasked stage and making acute challenge investigations more feasible.

Another obstacle to established experimental designs is the specificity expressed by some patients. That is, they claim to respond to only a limited number of specific chemicals. To conduct a controlled clinical experiment under such circumstances, challenge agents might have to be particular to each patient because of these individual variations. In single-subject designs, the subject of this paper, such an adjustment would be possible but would evoke other questions about the results, especially their extrapolation to other patients.

At this time, the primary question still asked by the biomedical community is whether MCS is an authentic diagnosis or camouflages some other condition such as depression. If the dominant issue is the validity of the MCS syndrome, the corresponding experimental question is whether it (and it is not yet unambiguously defined) can be elicited reliably with proper experimental controls. If this syndrome can be demonstrated, even in a restricted sample of patients, it can take its place as a valid diagnostic and even toxicological entity. Questions of prevalence will be deferred until the issue of validity is resolved. With so many unexplored dimensions remaining, the question represents prototypical exploratory research.

From this vantage point, we first seek to determine what is defined as internal validity; that is, how certain we are that manipulations of some independent variable, such as the ambient concentration of a specified chemical challenge, underlie variations in subject responses (the dependent variable). At this stage, we are much less interested in external validity, or the extrapolation of experimental results to other situations, groups, or environments. Sidman (*2*) stressed that the ultimate criterion of generality in science was not a statistical test and a *p* value, but replicability,

This paper is based on a presentation at the Conference on Experimental Approaches to Chemical Sensitivity held 20–22 September 1995 in Princeton, New Jersey. Manuscript received at EHP 6 March 1996; manuscript accepted 13 November 1996.

Preparation of this paper was supported in part by grant ES01247 from the National Institute of Environmental Health Sciences.

Address correspondence to Dr. B. Weiss, Department of Environmental Medicine, University of Rochester Medical Center, Rochester, NY 14642. Telephone: (716) 275-1736. Fax: (716) 256-2591. E-mail: weiss@envmed.rochester.edu

Abbreviations used: ADD, attention deficit disorder; AMA, American Medical Association; ARIMA, autoregressive integrated moving average; MCS, multiple chemical sensitivity; SD, standard deviation.

a criterion with special resonance for MCS. What we seek, then, is replicability across individuals at first and across settings (ecological validity) at some later time. This article proposes that research on MCS emphasize a class of experiments known as single-subject designs. It describes possible examples, then shifts focus to describe appropriate statistical procedures.

To set a context for this discussion, assume a normal distribution of sensitivity to specified environmental agents such that only a relatively minor proportion of the population responds adversely to current ambient levels. For example, assume that only those individuals beyond 2.5 standard deviations (SD) from the mean display sensitivity. Suppose that the ambient levels of these agents then shift slightly, as they might in a new environment such as a renovated building. Assume that under these new circumstances the distribution of sensitivity is displaced by 0.5 SD. Now, individuals beyond 2.0 SD fall into the sensitive category. Even a slight shift greatly expands the number of individuals in the sensitive zone because the consequences of such a shift are greatly magnified at the tails of the distribution. In the example above, the proportion of sensitive individuals rises from 0.62 to 2.28%. Such a phenomenon might account for why most Gulf War veterans did not have a problem. Those at the upper tail of the sensitivity distribution, who would have escaped many problems in their accustomed environments, found themselves exposed to higher levels of contaminants than they had experienced up to that time.

Faced with such a statistical conundrum, how does an experimenter proceed to select an appropriate research sample? The question of validity is not easily amenable to a search for characteristics that distinguish MCS patients from other groups because so many of those characteristics are self-defined. The syndrome is not cohesive enough to warrant group designs because such designs assume that differences between populations can be discerned by comparing their distributions on some outcome variable. But which populations would one compare? And under which circumstances? Suppose that an experimenter selects two groups for comparison. One is composed of MCS patients. The other is composed of nonpatient controls. How are they to be compared if they differ on multiple dimensions? Which criteria would be used for selection from the respective populations? Perhaps most crucial because the patient sample especially is certain to show appreciable heterogeneity, is which statistical models would be appropriate? A heterogeneous sample in which only a minor proportion consists of responders would require a dauntingly large number of subjects to demonstrate statistically acceptable differences because group averages obscure the responses of anomalous individuals (*3*).

A more persuasive source of data would be a longitudinal design in which patients are challenged repeatedly with both alleged triggers and inactive stimuli. Relationships established in even a few individuals would provide the basis for then exploring key variables and mechanisms. This is the design used to test what became known as the Feingold hypothesis. It is an instructive experiment because it embodies many of the same issues that confront MCS research.

The Feingold Experience

Feingold (*4*) claimed that many of the children whose behavior resembled that of those diagnosed with attention deficit disorder (ADD)–hyperactivity syndrome were simply exhibiting enhanced sensitivity to a variety of dietary elements. He included both natural constituents and additives. To test the hypothesis, we enrolled 22 parents whose childrens' behavior had been reported to improve with the imposition of a diet that eliminated the presumably offending diet constituents (*5*). While they were maintained on the diet, we intermittently challenged them with a blend of approved food colors at doses equivalent to our estimates of what was consumed daily by children between 3 and 7 years of age. During an 11-week period, we challenged them with such a blend, provided as a soft drink, on eight occasions. On all other days, they consumed a control drink indistinguishable in color and flavor from the color blend. Moreover, each child's response was gauged both with an individual set of 10 items culled by parents from a collection of behavioral inventories and by a standardized rating scale.

Two of the children displayed significant adverse responses to the food color challenge. The results from one of these children, a 34-month-old girl, appear in Figure 1. The chart compares her responses on several of the individual items and her scores on a standardized hyperactivity scale on control and challenge days. She clearly proved exceedingly sensitive to the food dye blend.

Experimental Designs

The approach taken for the color challenge study is a prototypical single-subject design. It has been used quite often in applied behavior analysis research for questions prompted by the success or inadequacy of various behavioral interventions (*6*). Behavior analysts find such designs attractive because of their origins in the experimental analysis of behavior and its emphasis on intensive study of individual organisms. Single-subject designs are also

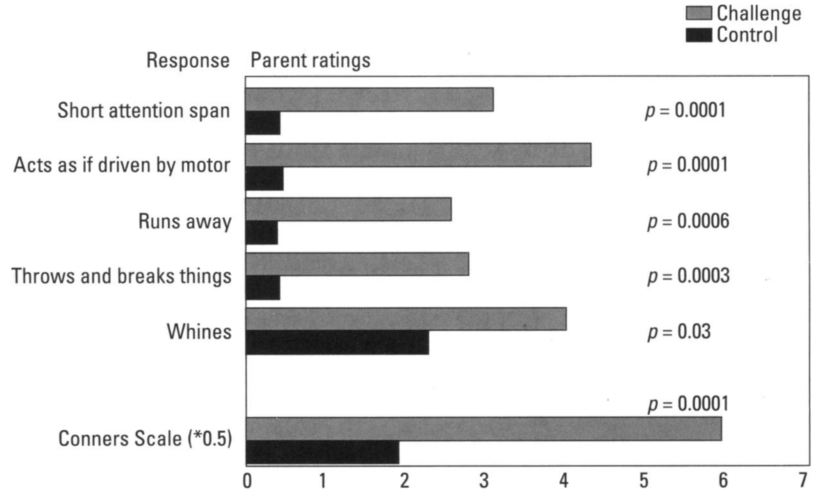

Figure 1. Parent ratings for a 34-month-old girl challenged with artificial food colors on eight occasions during an 11-week observation period (*5*). The subject consumed a soft drink on every day of the observation period. On randomly selected days, the drink contained a blend of food colors at doses based on dietary surveys of parents in the Kaiser system in California. On each day of the 11-week period, the parent recorded a rating score for each behavioral item, with the higher numbers designating a greater degree of expression. In addition, the parent also completed a standardized hyperkinesis scale. The *p* values are based on randomization tests (*11*).

encountered in other areas of research. Investigators who search for correlations between respiratory function and air pollutants (*7,8*) find them especially useful. Temporal correlations between ambient pollutant concentrations and functional indices, including diaries and symptom rating scales secured from individual respondents, provide the justification for positing a relationship. It might prove fruitful for some clinical studies of MCS to emulate those adopted by the air pollutant investigators if it were possible to measure an acceptable environmental exposure index.

For experimental MCS research, our special interest lies in interventions defined by programmed exposures. In such ventures, presuming that our primary medium is inhalation and that an exposure facility is available, we would enroll subjects who commit to an extended observation period so that reproducibility of response can be assessed. After an initial baseline period, they would appear at the facility at scheduled intervals, weekly or biweekly, for an exposure session during which various performance and other indices are recorded.

An adequate baseline period preceding any intervention is essential because introducing an experimental challenge may not produce an effect clearly indistinguishable from background. A subject's customary environment offers a variety of other agents and exposure sources, so customary variability in response measures is inevitable. The experimental schedule should be designed to determine if responses beyond this routine variability occur. Because our central question will be whether the experimental exposures provoke repeatable response patterns, how do we handle the possibility that exposure effects might persist beyond the exposure period? That question is intertwined with how we quantify effects.

As an example, suppose we focus on neuropsychological criteria; that is, performance tests and subjective state inventories. To capture subject status between programmed chamber exposures, we might emulate studies of the health impact of air pollutants in which subjects keep daily diaries (*8*). The diary entries are then correlated with area levels of specific pollutants. Structured diaries that also contain rating scales would provide useful sources of information about the impact of experimental chamber exposures beyond the exposure period. The food additive study cited above (*5*) asked parents to complete and mail data forms daily, and, with cooperative parents and an efficient staff, few data were lost.

For the particular aims of chamber studies, we could also take advantage of computerized testing. If we plan during exposure sessions to monitor functions such as memory, reaction time, and complex discriminative processes, and to gauge subjective state as well, an attractive option would be to equip subjects with inexpensive desktop or laptop computers. At a prescribed time each day, the subject would turn on the computer and proceed to load the program. He or she would then be presented with, for example, a 30-min test battery. Because forced-choice procedures are inherently resistant to sham-response patterns (*9,10*), they would also be included in the test collection. The results and all of the accessory information could be saved on a computer diskette and mailed, or transmitted by modem to the laboratory. An alternative procedure would use the modem to conduct the testing through a remote server. Using the security features of certain computer operating systems, it should be possible to control tampering with the tests or the results.

Types of Designs

For intervention designs such as chamber studies one possible alternative is called the ABAB design, as sketched in Figure 2. This design consists of a baseline or control treatment period of predetermined length, an intervention period during which some program of treatment is applied, another baseline or withdrawal period to contrast with the effects of the intervention period, and finally, a second intervention period. The second intervention period is essentially a reliability check of the first intervention period. The ABAB design is used for testing hypotheses based on a prolonged intervention regimen; for example, a 4-month experiment, with weekly observations, comprising a 1-month baseline, a 1-month intervention or exposure period, a 1-month withdrawal period, and a second 1-month intervention period. The ABAB design offers a firmer test of the intervention regimen than a variant called an ABA design because the second B period can be contrasted to the second A period despite a shift in the baseline. Because of its frequent application to experiments in animal behavior, it is often called an operant design.

ABAB designs might prove burdensome for questions about the efficacy of treatments for MCS because they are designed for lengthy interventions. A design that seems more suitable for experimental approaches to MCS imposes an intervention or experimental treatment on a quasi-stable baseline at specified times. A typical sequence might resemble that sketched in Figure 3, which depicts challenges or interventions of two kinds, control and active. Given the characteristics of the syndrome, particularly effects that linger beyond the exposure, we could provide adequate time between sessions to allow recovery or

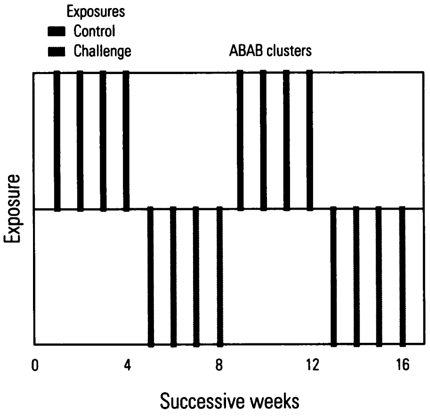

Figure 2. Schematic of the single-subject time series ABAB design. In this schematic, which depicts a 16-week clinical trial, exposure in a controlled setting takes place once weekly. The trial begins with 4 baseline weeks (A), followed by 4 exposure weeks (B), followed by a return to baseline period (A) and a second 4-week exposure sequence (B). In addition to data gathered during the exposure session, additional data are procured daily during the interval between exposures.

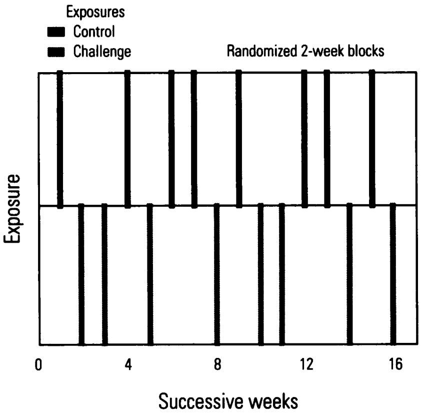

Figure 3. Schematic of a single-subject time-series design in which, for each pair of weeks during a 16-week trial, control or exposure conditions for the selected days are assigned randomly. Such a design is appropriate for analysis by randomization tests (*11*). A modified design alternates control and exposure weeks. As with the ABAB design, the period between experimental days can be used to acquire data on duration of effects beyond the exposure session as well as their character.

washout to baseline or to measure and trace the impact of the exposure.

For this example, the exposure schedule consists of a single weekly session. Separating consecutive chamber sessions by 1 week allows time for tracing the course of recovery to baseline. About half the sessions would be devoted to control exposures and half to experimental exposures in 2-week counterbalanced blocks, but the sequence within each block is chosen randomly, with each subject assigned an independently chosen sequence. Another option would alternate active and control weeks. Essentially, it is a longer ABAB sequence. This could have the advantage of offering more repeatable response patterns between exposure sessions and easier handling with statistical techniques designed to analyze periodicity. The same scheme is applicable to two different active treatments as well as to a comparison of control and active treatments.

Because the designs discussed above apply the intervention repeatedly, they permit the experimenter to determine whether a particular response pattern is reliably evoked by the intervention. The temporal aftermath of any single intervention may take different forms. The patterns depicted in Figure 4A–D can also describe responses to a more prolonged intervention: (*A*) the intervention produces a maximum response directly afterward that then fades; (*B*) the intervention produces a long-lasting, stable change; (*C*) the intervention produces a change that first rises to a maximum and then declines; and (*D*) the intervention produces a change that gradually builds. The possibility of a disrupted baseline in such a design must be acknowledged, and is one of the reasons for developing a monitoring scheme between intervention sessions. A disrupted baseline can be dealt with statistically in two ways. One is simply to wait until it returns, by some acceptable index, to stability. The other is to compensate for it mathematically, which can be accomplished by a variety of statistical procedures discussed later.

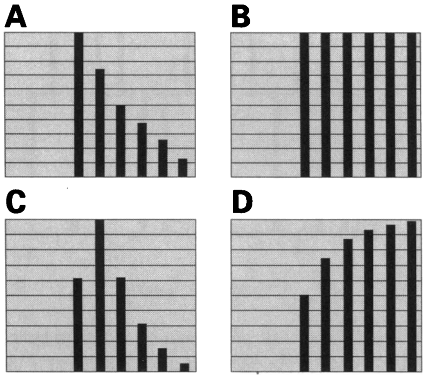

Figure 4. Hypothetical patterns of response following a single exposure to a challenge include: a marked elevation in response amplitude followed by a gradual decline (*A*); a significant response that remains at the same level (*B*); a gradual rise then gradual fall in response amplitude (*C*); and, a gradually rising response amplitude (*D*).

Analytical Procedures

Single-subject designs, by their nature, comprise a time series. Any sequence of time-ordered observations, in fact, comprises a time series. Fluctuations in stock market prices, historical temperature variations, and seasonal tracking of hospital admissions are all time series. Although the underlying process may be continuous, the observations themselves are spaced at constant intervals. The usual parametric statistical procedures typically applied to group designs may not be appropriate for single-subject experiments in which multiple observations are recorded over time to form a time series. Such experiments conflict with the underlying assumptions of parametric statistical models, namely, that repeated measures are independent and that error terms are uncorrelated. Repeated measures on a single individual are inevitably correlated, as are stock prices from week-to-week over an extended period. Neither varies randomly from day-to-day. Instead, they display significant serial or autocorrelations. Autocorrelated time series have the property that any single observation is predictable to some degree given the past behavior of the series. For most time series, the immediate past history is a better predictor than the more remote past history.

Many statistical procedures suitable for analyses of time series are available although uncommonly applied in research areas bearing on MCS. They are applicable to single-subject experiments because the data provide measurements scaled through time. Two analytical procedures, described below, take account in different ways of correlations among successive measurements based on the same subject. One, randomization tests, bypasses the underlying statistical structure of the time series in the interest of simplicity, although there are some design maneuvers to compensate (*11*). The other, time series analysis, explicitly incorporates autocorrelation.

Randomization Tests. The simplest method for analyzing a repeated intervention experiment such as the one depicted in Figure 3 is to compare the response to the experimental and control treatments by a randomization test (*11*). Randomization tests are distribution free; they make no assumptions about the underlying population, such as normality, and do not prescribe random sampling from such a population. They also make no assumptions about the statistical structure of a time series and are fairly simple to conduct. They are also fairly simple to interpret because, as Edgington (*11*) notes, they provide direct estimates of statistical significance without looking up information in tables or calculating probabilities defined by specific distributions. They could not be used in the past for any but the simplest experiments because of the enormous labor required to compute the permutations. It was only with the development of advanced computer technology that randomization tests became feasible.

In the experiment by Weiss et al. (*5*) described above, the 11-week period provided 77 scores for each outcome, 8 of which were associated with the food color blend. Each subject provided a mean obtained difference between color challenge and control drink days for each outcome. For each outcome criterion, 10,000 permutations (not a complete population, which would come to 77!/8!8!, but a sufficient Monte Carlo approximation) of the 77 scores were performed to provide a population of sequences. For each permutation, the mean of 8 randomly selected days was compared to the mean of the other 69 days. It was then possible to compute the proportion of permuted sequences that yielded as large a test difference as that obtained experimentally. This figure provided an exact *p* value.

With the scheme depicted in Figure 3, randomization tests would be applied as follows. Each 16-week period provides 8 control and 8 active interventions. These 16 scores are randomly permuted to yield an adequate sample population of 10,000 permutations. From each, 8 are chosen at random and their mean, for example, is compared to the mean of the remaining 8 scores. To provide a *p* value, we would then calculate how many such differences would exceed the experimentally determined difference. Levin et al. (*12*) as well as Edgington (*11*) describe how randomization tests can be applied to ABAB and similar designs in which each component could be a unit encompassing an extended period.

Time-Series Analysis. Successive observations taken in a longitudinal study such as those proposed for MCS research provide a complex statistical challenge. One source of complexity, noted above, arises from the intrinsic correlations between an observation at any particular time and observations at earlier times. Carryover effects from one occasion to the next, which induce trends in the data, provide another source. Because randomization tests, as noted earlier, make no assumptions about such statistical properties, they simplify analyses. Simplicity, of course, may not always yield the most desirable or interpretable result. Discarding complexities also means eschewing important mechanistic or process clues that might emerge if the data were carefully examined and analyzed. A more elegant, often more informative body of analytical techniques falls under the rubric of time series analysis. It takes as its task the interpretation of time-ordered observations of a process rather than an estimate of statistical significance. It may be especially useful for MCS experiments because it deals directly with the consequences of intervention.

The statistical apparatus for time-series analysis, although not generally familiar to scientists who rely on group designs, is mature and highly developed. Some of it represents a translation of the mathematical apparatus used for wave form analysis in engineering (*13*). Other components have developed from probability theory. Practical applications abound. Trends in stock market prices, for example, represent time series and have been the object of considerable statistical modeling. Trends in global temperature represent another, currently contentious time series because of the enormous implications attached to their interpretation. The air pollution literature referred to earlier relies heavily on time series analysis. Experimenters have applied time series analysis to uncover serial dependencies present in sequences of psychophysical responses (*14*) and of interresponse times in schedule-controlled operant behavior (*15*).

The basic model most often applied to time series of concern to us is denoted as the Autoregressive Integrated Moving Average model or ARIMA. One basic property of a time series described by such a model is that it is subjected to random shocks. Another is that the present state of the system exerts a greater influence on the system's output or succeeding state than any earlier states. The process that generates such a time series is described by the recurrence formula

$$x_t = f(x_{t-1}, x_{t-2}, \ldots x_{t-k}) + \varepsilon_t \qquad [1]$$

which signifies that the state of the system (for example, a subject's disposition to respond) at time t is a function of (is influenced by) his or her response dispositions at earlier time points. Here, ε is an error term taken to vary randomly. A special but common situation is described by the expression

$$x_{t+1} = ax_t + \varepsilon_{t+1} \qquad [2]$$

where $(-1 < a < 1)$ and ε again symbolizes a random shock. Equation 2 signifies a first-order autoregressive scheme, which represents a time series in which only the previous observation is needed to make the best prediction of the next observation. In this respect, it is also what is called a Markov process. The error terms are uncorrelated and are equivalent to white noise. The random shocks that account for the error term have been compared to an oscillating pendulum bombarded irregularly by small boys equipped with peashooters (*16*). Many, perhaps even most, of the time series we see in our research endeavors are primarily first order and can be characterized by Equation 2.

The primary statistical features of these time series, namely, the influence of prior observations, is important to grasp because they underlie the modeling applied to the actual data. Although the more commonly applied statistical procedures fit the data, ARIMA models do not rely on this method. Instead, the analyst determines the characteristics of the time series and then proceeds to build a particular model empirically. The modeling procedure begins with a consideration of trend, which McCleary and Hay (*17*) define as any systematic change in the level of the time series. If a trend exists, it must be removed because it violates the assumption of stationarity, that is, the property that the statistical characteristics of the time series are equivalent at all points in its history. Stationarity can also be violated if the time series drifts due to the accumulation or integration of random shocks over time.

Figure 5 shows schematically how such a process can occur. It reflects a situation in which some fraction of the response on one occasion (here it is 0.5) is incorporated into the amplitude of the response to the succeeding challenge. Its influence wanes with later challenges, falling successively by a factor of 0.5. Figure 6 shows how such influences can accumulate over successive occasions to produce an upward drift in response amplitude.

Removing the contribution of trend or drift to produce stationarity is usually achieved by a technique called differencing or, in wave-form analysis, prewhitening. It amounts to reformulating the time series as a succession of consecutive differences. The new time series is built by subtracting the first observation from the second, the second from the third, and so on. In most instances, the new series will now fluctuate about its mean value, the criterion of a stationarity time series . Stationarity is crucial if the model is to be used to define the effects of an intervention. If one differencing fails to produce stationarity, a second differencing is commenced.

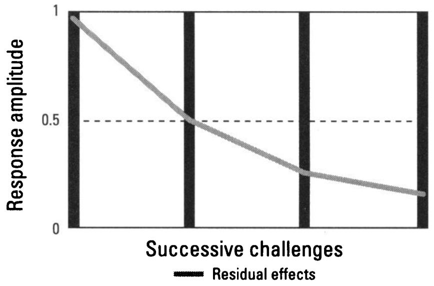

Figure 5. This schematic depicts the prototypical process in which the influence of any single observation persists but gradually wanes. In this diagram, the succeeding observation is influenced by a residual effect equal to 0.5 of the preceding observation's amplitude. Additional observations are influenced to lesser degrees; in the diagram, each succeeding observation declines by a factor of 0.5.

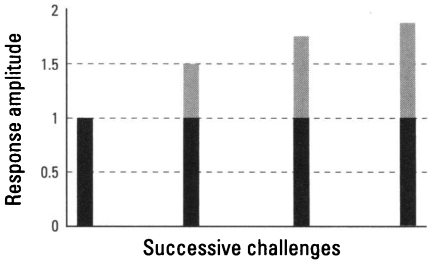

Figure 6. This diagram shows how serial dependencies extending over several points in a time series can accumulate to produce an upward drift in response amplitude. The technique of differencing (prewhitening) compensates for serial dependencies of this kind.

The autoregressive integrated moving average (ARIMA) model is described by three values: p, d and q:

- p represents the number of autoregression parameters; that is, the order of serial dependencies in the series. As noted above, most of the time series we deal with are typically first order ($p = 1$), meaning that only the previous observation and not earlier ones is significantly correlated with the present one. This is termed an autocorrelation of lag 1. If the serial dependencies extend to the two previous observations, then $p = 2$ (lag 2). An autocorrelogram is a plot of correlation magnitude over lag.
- d represents the number of differencing steps required to attain stationarity. In a situation in which $d = 1$, only one pass at successive differences (Observation 1 minus Observation 2, Observation 2 minus Observation 3, etc.) is needed to attain stationarity. Many times only this single step is necessary. If a second differencing is required, which is based on the new time series generated by the first differencing, the process is repeated.
- q represents the moving average component of the process. Moving averages are used to smooth the random component or noise in a time series and to forecast future values. In theory, q could be expressed as a weighted sum of past random shocks produced by autocorrelation among the effects of these shocks. Although in principle, a random shock can persist for q observations before its effect dissipates, in practice moving average models usually represent a process in which any single observation is a function only of the previous random shock. That is, q is usually equal to 1.

With the model able to account for serial interactions and trend, the experimenter is then able to ask whether including an intervention effect in the model enhances its predictive power. Put another way, the ARIMA model eliminates the contribution of serial dependencies to allow the intervention effect to emerge. Chamber exposure would be the event defined as intervention. Figure 4 depicted ways in which intervention effects can be modeled. By removing the components of the time series represented by the ARIMA model, the changes introduced by the intervention can then be determined.

To perform a time series analysis requires at least 40 repeated observations, however, so ARIMA modeling would not be appropriate for comparing 8 control to 8 active agent chamber days, as pointed out previously. For comparing chamber days alone, randomization tests would be appropriate. If chamber sessions took place weekly over an extended period, data acquired on all the intervening days could then serve as the basis for time series modeling. Because of the properties claimed for MCS, which imply some persistent effect arising from a chamber exposure, chamber exposures would then be defined as the interventions and independently evaluated by randomization tests. The aftermath of exposure would be modeled by time series analyses to determine a pattern.

Despite the difficulties posed by the kind of experimental design that would provide adequate data for ARIMA modeling, overcoming these difficulties is a worthwhile goal and not an insurmountable one. For the food additive study described earlier (*5*), parents were willing to contend with administering the soft drink at a prescribed time, completing several forms, making standardized behavioral observations, and contacting our staff every day for a 3-month period. MCS patients display no less persistence in pursuing the issues that concern them. Moreover, technology for remotely securing psychological test data is now available in a variety of forms.

Agents and Concentrations

MCS patients list such a broad array of triggers that the experimenter is confronted with the almost impossible task of choosing one or more that would be suitable for a challenge study of the type described earlier. Odor magnifies the problem of choice. Two alternatives are attractive. First, we should consider the possibility of using an agent such as ozone. Its virtues include a site of action in the deep lung, and, at reasonable concentrations, no sensory cues such as odor. Another virtue is the enormous experimental and epidemiological literature available on this agent.

An experiment incorporating ozone might proceed in the following way. On any single laboratory visit, the MCS subject would be assigned either ozone or a control exposure according to the designs portrayed in Figures 3 and 4. An expanded design would couple these conditions either with a distinctive odor such as that of amyl acetate or no odor. These four conditions would be repeated four times in different orders. Exposures would last at least 4 hr, but preferably 6 (*18*). A suitable concentration would have to be chosen. A value of 0.2 ppm will produce no persistent adverse effects in such an experiment.

During the exposure, the subject performs a variety of tasks. These should include prolonged vigilance and monitoring. In addition, exercise periods would be included. Both epidemiological and experimental studies indicate that subjects can become averse to exercise in the presence of relatively low concentrations of ozone or the ozone can induce mild deficits in pulmonary function (*18,19*). Although MCS patients do not typically cite ozone as a provocative agent, it seems to fit many requirements of the syndrome.

Another alternative is to expose the subject to an agent widely claimed to provoke reactions in MCS patients, but such a choice provokes difficult ethical questions. Volatile organic solvents are often mentioned as a class of agents likely to evoke symptoms in patients. Toluene, at the former TLV of 100 ppm (*20*) elicited performance decrements in healthy subjects during a 6-hr exposure sessions that also included 30-min bouts of exercise. Exercise on a bicycle ergometer raised both blood and breath levels of toluene. In a study with patients, exposure sessions with and without exercise designed essentially to produce differences in exposure, could serve as the contrasting conditions and still allow odor to remain as a variable.

These are only two model challenges among a broad sample of possibilities. They each display two assets. First, each is based on a body of experimental evidence demonstrating responses confined to acute effects in healthy subjects. Second, each is suitable for the kind of single-subject designs described earlier. The other alternative, mentioned earlier, is to choose, in consultation with the patient, an agent that he or she lists among the eliciting stimuli. The virtues of such an individualized approach are diminished by the obvious confounding it presents.

Interpreting Data

No matter how many testimonials and case studies are offered in support of the MCS syndrome, the biomedical community remains dubious about its existence, a response that only clear experimental data can reverse. But data are more than a vehicle for legitimacy. They are also a pool of information from which we extract hypotheses, mechanisms, and guides.

The conventional clinical trial is a group design. It often asks whether one treatment, administered to a sample of patients, is superior to another treatment or to a control treatment. If only a small subset of patients exhibits a significant positive response to the first treatment, its effectiveness will be concealed in group statistics. The single-subject design avoids this error but lacks the ability to generalize. At this stage of the MCS debate, the first aim of research should be to ascertain whether it can be demonstrated to exist, even in a circumscribed group of patients.

Would such a demonstration strengthen the argument that MCS is a meaningful diagnostic category? If only rare patients provide such a demonstration, MCS might be accepted as a valid phenomenon but not as a diagnosis. The experiment on food additives described earlier (*5*) disclosed two consistent responders. It did not lead to a diagnostic classification but demonstrated that, indeed, some children reacted with aberrant behavior to food dyes. Such a demonstration supported the contention that behavioral testing should be an important component of food additive toxicity evaluation. Should a consistent pattern of responses be demonstrable in MCS patients, such as one of those depicted in Figure 4, its validity as a diagnostic category would become far more cogent. Any consistent pattern, in fact, would provide not only an argument for diagnostic validity but even a means for making a diagnosis. Further, it would offer a model for exploring the biological substrates of the syndrome.

Some examples from the MCS literature illustrate how time series techniques might be used to clarify certain phenomena. Miller and Ashford (*21*) present a diagram to demonstrate the hypothetical impact on an individual's responses of fluctuations in environmental levels of multiple agents. The diagram depicts variations over time around what might be interpreted as a stable mean. Such a depiction is equivalent to a multivariate ARIMA model for which appropriate analytical tools are available (*17*). An accompanying chart presents another hypothetical time series that denotes responses to repeated challenges in an environmental unit. The chart diagrams a reliable response pattern, but given the inherent variability in response characteristics under such conditions, it would be hazardous to rely on the clinician's impressions as a guide to response validity. The authors even note that "At any particular time, how the person feels is determined not only by ongoing exposures, but by previous exposures whose effects may still be waning." Such a process, in fact, is shown in Figures 5 and 6 and is amenable to modeling.

Another example comes from the phenomenon of sensitization, defined as the progressive increase in sensitivity to repeated stimulations of various kinds. It has been suggested as a possible explanation of MCS. One variant of this process occurs in the form of kindling, which describes the increased probability of seizures with repeated chemical exposure or brain electrical stimulation (*22*). The data are typically presented in the form of group trends; in that mode, critical dimensions of the process are overlooked that might be ascertained by appropriate time series techniques. Serial correlations and other common features of time series are among those dimensions but uniformly ignored.

Finally, time series techniques are preeminent forecasting tools (*17*) and, in fact, widely used in health services research to predict hospital workloads, public health interventions, clinical test utilization, and many other indices critical to efficient planning (*23*). Although forecasting potential may not be an important attribute in an experimental context, it surely is a property that could prove useful in determining the contribution of certain interventions aimed at MCS. For example, an allied collection of symptoms, the sick building syndrome, has been investigated by modifying ambient conditions such as air flow, humidity, and temperature. An appropriate question to ask about such interventions is the extent to which they produce sustained effects once they have commenced.

Most of the widely adopted statistical software packages such as SAS, BMDP, SPSS, Minitab, and others, contain routines for conducting time series analyses. Naturally, it would be wise to consult with a biostatistician before embarking on such an analysis, but few contemporary efforts in biomedical research can be accomplished without the contribution of several different kinds of specialists.

Epilogue

This paper asserts that the most convincing source of experimental data from chamber studies is the single-subject design. Instead of examining deviations from a group mean, it focuses on individual patterns of response. It maintains the uniqueness of the subject, a crucial factor in MCS research because of the continuing debate over case criteria. It proposes, as an alternative to common group designs, repeated observations in individual subjects that permit the application of two statistical techniques. One, randomization tests, allows a direct estimate of experimental versus control differences on appropriate outcomes. The other, formal time series analysis, allows the time series itself to be modeled. Because this technique requires long series of observations, it would require a design in which weekly chamber sessions, for example, would be treated as interventions; on all other days, observations would be treated as components of the time series. Although much more complex, time series analysis offers an opportunity to examine in detail the impact of an experimental exposure.

REFERENCES

1. Council on Scientific Affairs, AMA. Clinical ecology. JAMA 268:3465–3467 (1992).
2. Sidman M. Tactics of Scientific Research. New York:Basic Books, 1960.
3. Weiss B. Low-level chemical sensitivity: a perspective from behavioral toxicology. Toxicol Ind Health 10:606–617 (1994).
4. Feingold BF. Why Your Child Is Hyperactive. New York:Random House, 1975.
5. Weiss B, Williams JH, Margen S, Abrams B, Caan B, Citron LJ, Cox C, McKibben J, Ogar D, Schultz S. Behavioral responses to artificial food colors. Science 207:1487–1489 (1980).
6. Kratochwill TR, ed. Single Subject Research. Strategies for Evaluating Change. New York:Academic Press, 1978.
7. Neas LM, Dockery DW, Koutrakis P, Tollerud DJ, Speizer FE. The association of ambient air pollution with twice daily peak expiratory flow rate measurements in children. Am J Epidemiol 141:111–122 (1995).
8. Schwartz J, Wypij D, Dockery D, Ware J, Zeger S, Spengler J, Ferris B. Daily diaries of respiratory symptoms and air pollution: methodological issues and results. Environ Health Perspect 90:181–187 (1991).
9. Pankratz L. A new technique for the assessment and modification of feigned memory deficit. Percept Motor Skills 57:367–372 (1983).

10. Heaton RK, Smith HH, Lehman RAW, Vogt AJ. Prospects for faking believable deficits on neuropsychological testing. J Consult Clin Psychol 46:892–900 (1978).
11. Edgington E. Randomization Tests. New York:Marcel Dekker, 1987.
12. Levin JR, Marascuilo LA, Hubert LJ. Nonparametric randomization tests. In: Single Subject Research. Strategies for Evaluating Change (Kratochwill TR, ed). New York:Academic Press, 1978;167–196.
13. Gregson RA. Time Series in Psychology. Hillsdale, NJ:Erlbaum, 1983.
14. Weiss BW, Coleman PD, Green RF. A stochastic model for time ordered dependencies in continuous scale repetitive judgements. J Exp Psychol 50:237–244 (1955).
15. Weiss B, Laties VG, Siegel L, Goldstein D. A computer analysis of serial interactions in spaced responding. J Exp Anal Behav 9:619–625 (1966).
16. Bartlett MS. Stochastic Processes. Chapel Hill, NC:Institute of Statistics, University of North Carolina, 1947.
17. McCleary R, Hay RA. Applied Time Series Analysis for the Social Sciences. Beverly Hills, CA:Sage Publications, 1980.
18. Weiss B. Behavior as an endpoint for inhaled toxicants. In: Concepts in Inhalation Toxicology (McClelland RO, Henderson RF, eds). Washington:Hemisphere Publishing, 1989;493–513.
19. Weiss B, Rahill AA. Applications of behavioral measures to inhalation toxicology. In: Concepts in Inhalation Toxicology. 2nd ed (McClellan RO, Henderson RF, eds). Washington: Taylor and Francis, 1995;505–532.
20. Rahill AA, Weiss B, Morrow PE, Frampton MW, Cox C, Gibb R, Gelein R, Spears D, Utell MJ. Human performance during exposure to toluene. Aviat Space Environ Med 67:640–647 (1996).
21. Miller CS, Ashford NA. Chemical Exposures. Low Levels and High Stakes. New York:Van Nostrand Reinhold, 1991.
22. Weiss SRB, Post RM. Caveats in the use of the kindling model of affective disorders. Toxicol Ind Health 10:421–447 (1994).
23. Weiss TW, Ashton CM, Wray NP. Forecasting areawide hospital utilization: a comparison of five univariate time series techinques. Health Serv Manag Res 6:178–190 (1993).