SOME UNSOLVED MODEL BUILDING PROBLEMS IN APPLIED BIOSTATISTICS[1]

by

Dolores S. Smith and David S. Salsburg[2]

I: The Analysis of Data from Psychotherapeutic Drug Screens

Developments in the conduct of experiments in animal psychology have in many ways gotten well beyond the current state of statistical sophistication. Studies that appear in the literature often present a statistical analysis of two-dimensional contingency tables derived from the data. Results are invariably accompanied by a statement of the sample size and the level of significance. Contingency table tests serve to satisfy an elementary editorial requirement that something was "statistically significant". However a closer look will reveal that major conclusions are based on eyeballing the complex pattern of results which the experiment shows.

A typical experiment of a drug screen is described as follows: An animal is premedicated with the experimental compound, then given a sub-lethal dose of a challenge drug of known effect. The known effect of the challenge drug consists of a sequence of stereotyped patterns of animal behaviour. For instance, strychnine will induce first escape activities, then erratic random movements, then convulsions, and finally death. If the dose of the challenge drug is high enough, the probability that the animal will go through the entire sequence is nearly 1.0. Such a high dose of the challenge will also block all but the most potent drug screens. The

[1] Paper presented at the Statistical Meeting in India, Dec. 1977.

[2] Dolores S. Smith is a faculty member of the Dept. of Statistics, University of Connecticut, U.S.A.; Dr. David S. Salsburg is Senior Statistician of Pfizer Central Research, U.S.A.

aim of the initial screen is often to identify a class of compounds with some activity in order to work up increased potency. For this reason the dose of the challenge is adjusted in such a way that the probability of final death runs from .5 to .8, and is a random function of day to day differences in laboratory conditions. (Sometimes there is an initial run-in on a given day or at the beginning of the week to determine the dose of the challenge.)

For each animal the time to each point of stereotyped behaviour is recorded. The behaviour sequence is so predictable as an animal crosses from one behaviour pattern to another. Screens currently used by many drug companies involve replicating from 5 to 10 animals on a given experimental medication. There may be up to five such experimental compounds on a given day, compared to two controls, the controls being no pre-medication and an active drug of known effect.

Most of the collected data go unanalyzed. The active control group is often used only to show that the screen exhibited a drug effect on that day. (The entire day's data is thrown out if this is not so.) Often an endpoint short of death is chosen that has a high probability of occurrence in the controls, e.g., convulsion in the strychnine challenge. A nonparametric test is used to compare the time to that event among the groups. It is expected that different kinds of drugs will lengthen or shorten the intervals of time to the endpoint. Once a compound is found "significant" on a statistical test, it is common for the experimenter to "eyeball" the pattern of mean times between events to characterize the drug in terms of its qualitative activity.

Most drug companies have data banks involving hundreds of thousands of animals. Data against which to check or model hypotheses are available in machine readable form.

Initial investigations by statisticians at one pharmaceutical company suggest that time to a specific endpoint does not fit an exponential distribution, but seems to fit a Weibull distribution. However both parameters of the Weibull distribution change from time to time and from drug to drug.

Problem: Is it possible to specify a model for the vector of intra-event times that will yield reasonably powerful statistical tests even for small samples, a model that is sufficiently flexible to take into account day-to-day changes in the experimental set-up and at the same time yield easily understood parameters that describe specific drug classes? In particular, can a Bayesian model be constructed that utilizes the vast files of animal data that now exist?

II: Human Challenge Studies

Challenge trials in the clinical evaluation of new drugs are based on a deterministic pharmacological model which works well with animals but cannot be used in its deterministic form in humans. A prototypical pharmacological challenge study is the guinea pig anti-histamine "screen". A guinea pig reacts to a sufficiently high level dose of histamine in a stereotyped way: first it begins to pant and breathe with difficulty, then it races around the cage, then goes into convulsions, then rolls over and finally dies. A guinea pig is pre-medicated with a drug, exposed to a dose of histamine titrated at the beginning of the day's experiment, and observed. Too low a dose of the challenge will fail to elicit the stereotyped behaviour. Too high a dose will not be blocked by the premedication. So a range of histamine doses is determined which will evoke the stereotyped behaviour in unmedicated animals and be blocked by known anti-histamines. This range is usually quite broad and the genetic uniformity of the animals guarantees that the response will be reproducible.

The human version of challenge studies differs in several respects. Humans differ in their responses to a fixed challenge both among themselves and across time within themselves. There is risk involved in administering too high a dose of the challenge. Thus individual subjects are subjected to a slowly increasing dose of the challenge until it just barely elicits a specific response. This will determine the maximum dose. On successive days the patient is pre-medicated and exposed to the maximum dose of the challenge or slightly beyond. The end point is determined when an evoked response is observed.

The allergen challenge for prophylactic treatment of extrinsic asthma is typical. Through skin testing the allergens to which a patient is sensitive are found. A mixture of allergens is produced, measured in units of equivalent response (PNU) using certain linear formula developed many years ago. On titration day the patient is exposed to an increasing number of PNU units of allergen. These are measured in cumulative number of units since it is believed that previous doses have residual effects that are added onto the succeeding doses. After each dose the patient is tested on a spirometer. A response occurs when the patient goes into a broncho-spasm or shows a deterioration on the spirometer measure.

Thereafter the patient is pre-medicated and then given a sequence of doses of the challenge. Spirometer measures are taken after each challenge. The general medical approach is to consider a patient as "protected" if he does not deteriorate at the maximum challenge level that caused deterioration on titration day.

Almost all of the accumulated intermediate data are ignored. If anything is quantified, it is usually the cumulative PNU value that leads to final deterioration. Unfortunately there is considerable random noise in the experiment. Patients may be "protected" at a given dose but not at a higher dose, or they may be "protected" at a certain dose of a drug but not "protected" when given a repeated dose of the same drug. Patients tend to accommodate to the continuing use of the same level of allergen, some becoming more and more sensitive, while others becoming more and more resistant so that they tend to be "protected" by placebo with increasing probability.

More complications arise. Spirometer measures appear to have discrete distributions. For any patient on a particular day all measurements land on 4 to 6 unique values. In the absence of a challenge or medication the distribution appears to be symmetric with 50% or more concentrated on one value and 20% falling on two adjacent values. Unfortunately, it is not possible to anticipate the unique values for a given day. They change from day to day and from patient to patient. Since the challenges are fixed, the cumulative PNU values leading to deterioration are also taken from a discrete distribution.

Another challenge study deals with exercise testing of patients with angina. There is a method of exercise challenge which has been fixed and used universally so that drugs and experiments can be compared. It is called the Bruce treadmill test. The patient is set to walking against a powered treadmill on an incline. The speed and incline of the treadmill are increased at fixed points in time, with the pattern of increases fixed in the protocol. The pattern is such that each new step represents a quantum jump in exertion. Only superbly conditioned athletes are able to get through the entire sequence without something occurring.

The "something" that occurs determines the endpoint of the challenge. The typical protocol will have the patients continue on challenge until angina pains become "unbearable", a "significant" cardiac event occurs (like tachycardia), the patient complains of extreme fatigue, or, in the clinical judgment of the physician, the patient has had enough. The exact stage at which this occurs is recorded. Measures of cardiac parameters are taken, such as blood pressure, pulse, an EKG tracing, and blood gases.

The usual drug trial has a patient run through exercise testing without drug or with one single-blind placebo one or more times for training and to establish baselines. Then the patient is given medication, usually a chronic therapy for several weeks, and brought in for an exercise test. Usually the study then crosses the patient over to another medication for a similar test. Frequently a patient will stop on the treadmill for different reasons at different times.

These challenge studies have similar problems:

(1) all observations stop as a result of a random event which is part of the sequential operations

(2) one of the measures is the endpoint of a discrete titration of challenge

(3) multivariate observations are taken at each point in the challenge, but they are only weakly related to the endpoint, although one or more of them might be used to define the endpoint.

Problem: Can the time course of the challenge be modeled? Can "baseline" information be utilized analogous to the use of covariates in linear hypothesis testing? Can the discreteness of the observations be used to construct a contingency table onto which mathematical models can be imposed? A typical study utilizes from 20 to 30 patients (replicates). Thus any method of statistical analysis will belong to the "moderate" sample size category and should not lean too heavily on asymptotic results.

III: In Utero Exposure to Suspected Carcinogens

The concept of *in utero* exposure has been around in carcinogensis work for some time for two reasons: Firstly, it provides additional impetus to the testing of a carcinogen. It is not possible to detect events of low probability with a small animal study, but we can increase the probability of tumor attacking the animal *in utero* before it develops its immune mechanisms. Secondly, many ubiquitous substances, such as food additives, food coloring, etc., are potential carcinogens and might be ingested by pregnant women. This possibility should be modeled in animal studies.

At the conclusion of the study there will be data from two generations, F_0, the parent generation that was fed the suspect material since weaning and F_1, the offspring generation that will have been exposed since conception. From 18 to 30 organs will be examined and each marked zero or one to indicate absence or presence of a tumor. Hence observations from a single animal consist of a vector of zero's or one's.

Actually the data structure is more complicated. One can identify several stages of the lesion in question, from mild hyperplasia to benign tumor, to malignant tumor, to invasive or metastatic tumor. Some test animals will have died during the study; others will have been sacrificed while still "healthy" at predetermined points of time. If the animal is prone to have tumors in a specific organ, then the early appearance of tumors in the treated group indicates an effect. The identification of a specific tumor type is subject to considerable disagreement among pathologists and within the same pathologist faced with the same slide at different points in time. Some animals will have died with partial autolysis and some elements of their vectors will be missing. It is standard practice to take a slice of tissue for histopathological examination but additional specimens are taken if the animal appears to have gross lumps that the prosector might deem suspicious.

Let us assume that the individual datum consists of a vector of zero's and one's. The F_0 generation will have been assigned at random to treatment and controls but there is serious question about the nature of the F_1 generation. The protocol might call for a fixed number of males and females to be retained out of each litter. If the experimental compound tends to increase fetal wastage, the litters in the test group will be smaller. It is animal handling practice to cull out the weak pups and retain only the vigorous ones. Animal handlers know from experience that a substantial number of less vigorous pups will not survive weaning or early handling. When the litter is small in size, it may be necessary to retain less vigorous pups in order to meet the quota. This can be controlled to some extent by increasing the number of F_0 females and retaining the "best" of the litters, but then the choice of pups or litters is not random.

Thus, in any practical situation, the F_1 generation consists of litters whose numbers are the maximum of a random number and the fixed quota. The probability of an event for an animal within a litter is a function of the treatment and the genetic component inherited from the parents. If the causes of many tumors are latent viruses, say from infections given by the mother or by littermates before weaning, some sort of nested effect will be independent of the father's genetic component.

At present the standard method of statistical analysis is to consider the total number of animals in both generations with a given tumor type and run a 2 x 2 contingency table analysis of the counts between treated and controls. This was done in the Canadian study which showed that saccharin "caused" bladder tumor in male animals. There has to be a better way.

Problem: The vectors of zero's and one's have a correlation structure that can be estimated from historical controls, but it appears that active compounds will also shift the correlation structure. Any statistical test that depends on the estimated control correlation structure should be sensitive to changes in both the first and second moments of the vector. Can this sort of vector be modeled? Can the components of probability be modeled in a way that will lead to internal verification from the data? Can a model be constructed which will enable one to compute the operating characteristics of specific protocol designs, e.g., the number of animals per sex retained in each litter?

This problem has relevance and a certain amount of urgency at the moment. Under the impact of the saccharin studies, the food industry of the U.S. and Canada is beginning a massive series of *in utero* tests of common food additives. The first of these tests will cost over $4 million and was started in June 1977. They are bound to lead to some ambiguous events, such as the occurrence of tumors in almost all animals of some litters. Can an unbiased and powerful method of hypothesis testing be evolved before they come "off the boards" in late 1979? A similar study on a small scale is in progress now. The National Cancer Institute is expected to issue a report concerning the results of some tests linking a dry cleaning chemical (perchloroethylene) to liver cancer in mice. Both chemical and drycleaning industries are expected to issue reports of their own inhalation studies to disprove the claim that the chemical is a cancer hazard to people. As in the aforementioned studies powerful methods of hypothesis testing are needed before general statements are made public.