

Parallels between Statistical Issues in Medical and Meteorological Experimentation

K. RUBEN GABRIEL

Department of Mathematics, University of Rochester, Rochester, New York

(Manuscript received 19 May 1998, in final form 6 May 1999)

ABSTRACT

The methodology of experimentation, randomization, and statistical analysis in weather modification has many parallels in clinical trials, such as the need for randomization, and the question of inclusion or exclusion of units assigned to be treated but not actually treated. There also are considerable differences, mainly in the definition of units, where the obvious choice of a single patient is in contrast with the highly problematic definition of a cloud or storm, and in the ethical aspects. This paper highlights some of these parallels and differences in the hope that looking at one's own problems in a different context may enhance one's understanding. It may also reconcile experimenters to their need for statistics: as the Hebrew saying goes, "*Tzarat rabim, hatzi nehama*" (the misfortune of many is half a consolation).

DEDICATION

It has been my good fortune to have worked with meteorologists who had an appreciation of what statistics could do for them and even seemed to get pleasure from understanding it. Such was Graeme Mather: not only did he seek and heed statistical support for his seminal work in cloud seeding, but he conveyed a sense of enjoying the cooperation and intellectual exchange. When we collaborated, I was challenged to propose and to justify new methods of analysis—as in exploring Nelspruit rainfall by means of biplots and linear models (Gabriel and Mather 1986) and in a joint attempt, by e-mail, to apply QQ-plots to the study of detailed differences between the rain distributions of seeded and unseeded storms—and I was gratified to see his intelligent and useful application of these analyses. Graeme not only made his own important contributions but also had a singular gift of making others feel that he understood and appreciated them. It is appropriate to dedicate the following paper to him, not only because I have learned much of what I write about from collaborating with him and other like-minded meteorologists, but also because he was present when I delivered the initial version in Bari, Italy, in 1996 and was generous in expressing his appreciation. I miss him.

1. Introduction

Weather modification experiments take years to complete; involve use of statistical designs that include controls, randomization, and covariates; and are summarized by statistical analyses. Yet such experiments have not generally established clear treatment effects despite the cloud seeders' frequent conviction that they had affected clouds and precipitation. This failure has tempted the meteorologists to blame the statisticians: How could *their* calculations miss the dramatic results of *our* cloud seeding?

There are parallel experiences in medical research, as shown by a review of 53 published surgical studies,

each of which compared an innovative procedure with a standard. Table 1 shows that among the well-designed trials none of the results were encouraging whereas among the poorly designed trials most results appeared to be encouraging. Does it follow that careful experimentation impedes good surgery, or does it suggest that many published claims arise from poor experimentation rather than from good medicine? The parallel with weather modification is striking, as it would be for any field in which treatment effects must be evaluated against a background of random variability. It is easy to produce "encouraging findings" by cutting corners in scientific rigor.

Methodological and statistical issues in research have parallels in many different fields, and the solutions proposed in one are often instructive for workers in others. Meteorologists may find it interesting to compare their concerns with those of medical scientists and vice versa. This paper encourages such comparison by illustrating

Corresponding author address: K. Ruben Gabriel, Department of Mathematics, University of Rochester, Rochester, NY 14627.
E-mail: krg1@troi.cc.rochester.edu.

TABLE 1. Degree of control vs degree of investigator enthusiasm in 53 surgery studies.

Experimental design	Degree of investigator enthusiasm about results			Total
	Marked	Moderate	None	
Controls, randomized	0	3	3	6
Controls, not randomized	10	3	2	15
No controls, not randomized	24	7	1	32
Total	34	13	6	53

Source: Gilbert et al. (1977).

parallels in experimental design, bias, randomization, and statistical analysis in cloud seeding experiments and in medical experiments on patients, usually referred to as “clinical trials.” The examples that are presented are very simple, but, except for one hypothetical example, they are all real.

2. Some concerns of ethics and of science

Before comparing the techniques of experimentation and analysis in weather modification and in medical treatment one must consider the ethical issues involved, both in the experimentation itself and in applying its results.

Experimenting with human subjects is obviously different from experimenting with clouds or storms. Individual rights that must be safeguarded in the former have no immediate counterpart in the latter, except in so far as concerns the effect of cloud seeding on people if it creates hail, or floods, or other adverse effects. The requirement of informed consent is accepted as the patients’ right but there are no parallel rights for the populations potentially affected by weather modification, though there have been a few cases of litigation.

Application of the results of experiments, on the other hand, has close parallels between medical treatment and weather modification. It is just as much an ethical imperative to apply knowledge gained from a weather experiment in a policy to alleviate a drought, as it is to apply knowledge from clinical trials to a practice that reduces patients’ suffering. The emphasis on the one or the other presumably has to do with one culture focusing more on the community and another on the individual. While Thailand has instituted a Royal Rainmaking Bureau that attempts to relieve food shortages for the population at large, the United States has put far greater emphasis on supporting the medical establishment’s attempts to cure patients as individuals.

Issues of ethics are also connected with the phase of experimentation. The ethical issues are much more acute when the experimental results have immediate applicability than when the results are a relatively early step in building up an understanding whose practical application may be feasible only at a much later time. Much weather experimentation is concerned with basic sci-

ence: What effect does introduction of a glaciogenic or hygroscopic agent have on any feature of the clouds? Many clinical trials, on the other hand, are focused on finding out whether a drug, or intervention, will have a specific effect, usually on survival. The findings of clinical trials often translate directly into decisions on treatment of future patients. The findings of cloud seeding experiments, on the other hand, do not translate simply into weather modification policies. It is a long way from an experiment’s showing that cloud seeding augments radar reflectivity in a cloud to gauging the economic benefit of a seeding operation, or even to knowing whether it would produce more precipitation to the ground.

Of course, some medical experiments also have the characteristics of basic science, such as pharmacokinetic studies and bioequivalence trials, and some meteorological experiments are testing the practical application of a technique, such as those of fog dispersal at airports and of reduction of hail damage. By and large, however, clinical trials and weather modification experiments differ in the phase of acquiring knowledge and application to which they are directed. That is one reason why they differ in the urgency of the ethical issues they raise.

3. Experimental units and treatment allocation

a. The nature of experimentation

The nature of an experiment is manipulation with a view to observing whether an effect results. That process involves experimental units, treatments that are applied to them, responses that are observed, and methods by which the effect is evaluated.

Some experiments are by trial and error, adapting a search for suitable treatments in promising directions, whereas other experiments are designed and executed according to a rigorous protocol that permits confirmation or rejection of an a priori hypothesis. In the study of drugs, what is known as phase I, and possibly phase II, consists of adaptive experiments that “check safety and establish a dose at which studies in men may be undertaken,” whereas phase-III experiments are designed for “proving efficacy to a sceptical regulator” (Senn 1997, p. 2). In weather modification, adaptive experiments with instruments and burners and observation of individual cloud’s apparent reactions precede experiments that are designed to confirm the efficacy of a seeding technique in a given time and place.

Adaptive experimentation is a learning experience that usually cannot be evaluated objectively. Designed experiments, by contrast, are tightly controlled “black boxes” with a protocol that is planned ahead of time and continued unchanged to completion, when the results are analyzed by already planned statistical methods, including significance tests. [See also Mielke (1995), who refers to the two types of experiments as

exploratory and confirmatory, terms that I prefer to apply only to the analyses.]

The distinction between adaptive and designed experiments is crucial. The current discussion by a statistician is mostly about the latter and stresses the need for rigor in their planning, execution, and analysis. There is no doubt that designed experiments are of paramount importance in testing medical treatments before sanctioning their use, and the rigor of their design and execution is insisted on by the regulatory agencies. Their appropriate role in current weather research is much less clear and is monitored by peer review. Is it time for designed experiments to test the effectiveness of techniques of cloud seeding? Would it be better to devote more effort to adaptive experiments that search for any meteorological signatures of various advertent and inadvertent influences? This paper can only refer to that issue, not resolve it, but it must point out that it is not appropriate either to apply the rigor of designed experimentation to adaptive probes or to ignore statistical rigor in an experiment designed to test a hypothesis.

b. Definition of experimental units

The basic unit used for experimentation is much easier to define in medicine, where it usually is an individual, than in meteorology, where it may be something as elusive as a cloud. Kempthorne (1980), comparing research on the weather with agricultural research, says that in the former

the experimental unit is, at best, a combination of a time point or brief interval with a fuzzy “pig” flying in the sky. The “pig” is not at all well-defined as regards boundaries. It has a very obscure and uncertain life over time. It is here as a definite, perhaps, region of space at time t_0 ; at time $t_0 + x$, it is somewhere else in space; we may not even be able to identify it as being the after-cursor (c.f., pre-cursor) of the cloud we *think* we have seeded. How can we possibly claim to have an analog of our real pig experiment. . . ?

Weather experimenters must spend much more effort on objective definition of their experimental units than must medical researchers. Basically, they have a choice of either providing an operational definition based on a sophisticated evaluation of radar echoes—as, for example, in the experiments in Texas (Rosenfeld and Woodley 1989), Illinois (Changnon et al. 1995), South Africa (Mather et al. 1997), Thailand (Silverman et al. 1999; Woodley et al. 1999), and Mexico (Bruintjes et al. 1999)—or of rough and ready simplifications such as the precipitation on rain gauges in a well-defined target during a well-defined period such as a 24-h day—as, for example, in Israel (Gabriel 1967) or Tasmania (Miller et al. 1979).

c. Inclusion and exclusion of units

Clinical trials usually address the efficacy of a treatment under fairly well specified conditions. Their protocols therefore include rigid criteria for inclusion and exclusion. A study of postoperative thrombosis, for example, “excluded [patients] if they were under 18 years of age, were having a revision of a previous (surgery), were allergic to [etc., etc.] . . .” (Francis et al. 1992). The well-known randomized study of the effect of aspirin on heart disease (Steering Committee of the Physicians’ Health Study 1989), considered only men; that was straightforward, though probably unfair to women about whom the study provided no information. It often is not obvious who is a (potential) patient in a particular medical context. Nor is it self-evident what meteorological unit is most relevant to obtaining a desired weather modification. One needs to balance the attempt to narrow down the study to a relatively homogeneous sample against the hope that the study’s conclusions will be widely applicable.

As in clinical trials, weather experiments require a clear protocol as to which potential units are to be “enrolled” in the experiment. Thus, the randomized Grossversuch III on hail prevention experimented only on days with favorable hail forecasts (Schmid 1967). The 1989 Illinois experiment is an instance in which elaborate criteria were laid down to decide that a cloud was to be included if “it had a) a cloud top just passing 20,000 feet, with potential for reaching 30,000 feet and beyond; b) a cumulus congestus (hard, blocky) appearance; c) [to] be at least 2 kilometers in diameter; and d) little or no vertical tilt” (Changnon et al. 1993).

Actual inclusion, subjection to treatment as planned, and continuation to follow up observation, is not quite the same as what is defined in the protocol. In the postoperative experiment above, for example, some enrolled patients declined to participate or withdrew before surgery, and others withdrew or were otherwise lost to the study after having been allocated a treatment or having started with it. Excluding the former cases was not going to bias the analysis, but excluding the latter might have biased it if one of the treatments would have influenced more patients to withdraw than the other. In general, there is a problem of whether the analysis should be of all those “intended to be treated” or only of those treated and observed “per protocol.” The biases that can arise from excluding the “noncompliers” depend on whether the treatment might have affected noncompliance with the protocol [Senn 1997, chapter 11; see also Oakes et al. (1993) for a study examining the effects of noncompliance on possible biases].

The parallel problem in weather modification occurs in the temptation to exclude units assigned to be seeded if they present no suitable clouds. The effect of such a procedure can be illustrated by the alternating target estimate (Gabriel 1999):

$$\frac{\text{precipitation on target when seeding allocated to target itself}}{\text{precipitation on target when seeding allocated to other target}} - 1,$$

which was used, with various adjustments, to estimate seeding effects on two targets seeded alternately. In Puglia, this estimate was found to be negative and close to significance for experimental days on which seeding was not carried out (List et al. 1999). Because there could be no effect of treatment without seeding, this negative estimate was evidence of a selection bias that is likely to have occurred in the following manner. Seeding was not carried out when there were no clouds on the allotted target, even if there were clouds on the other target. No wonder those days produced a negative estimate of “seeding effect.” Conversely, a positive bias must have occurred on seeded days, because such seeding was carried out only when there were clouds on the allotted target, even if there were none on the other target.

d. More on selection bias: Examples

Limiting analyses to units actually treated, however intuitively appealing, is a source of positive bias and must be avoided. This type of bias is subtle enough to confound many experimenters and may be worth belaboring by means of the following imaginary “medical experiment.” Imagine pairs of fraternal twins and random selection of one twin from each pair for a treatment intended to increase height; the response Y that is recorded after treatment is how much taller the treated twin is than the untreated twin. If the treatment were ineffective, the expected (i.e., overall average) difference $E(Y)$ would be zero, whereas $E(Y)$ would be positive if the treatment were effective. Three experimental designs are considered: D1 to use all twin pairs, irrespective of sex; D2 to exclude all pairs in which the selected twin was female; and D3 to exclude all pairs in which either twin was female, that is, to include only twin males.

For simplicity, assume all males to be 175 cm tall and all females 165 cm, and let the proportion in each sex be 1/2 and the two twins’ sexes be independent (as they are in fraternal twins). Table 2 then gives the distribution of the twin pairs, their heights, and the differences Y under the hypothesis of no treatment effect. It also shows which types of twins are included under each of the three designs and with what probabilities. Last, it shows the value of the expected height difference $E(Y)$ under each of the designs. This is simply the average of the Y s over the types of twin pairs included.

Experiments using designs D1 or D3 are unbiased because they have $E(Y) = 0$, which is the correct mean under the null hypothesis; an experiment with design D2 will be biased because it has $E(Y) = 5$, which is

TABLE 2. The expectation of a variable under different experimental designs.

Pair of twins					Experimental design		
Selected twin		Other twin		Height difference Y	Probability of choosing such a pair		
Sex	Height	Sex	Height		D1	D2	D3
M	175	M	175	0	0.25	0.50	1.00
M	175	F	165	10	0.25	0.50	0
F	165	M	175	-10	0.25	0	0
F	165	F	165	0	0.25	0	0
Expected mean height difference $E(Y)$					0	+5	0

wrong. The reason for the bias of design D2 is that its selection of twin pairs depends on a covariate (the sex of the treated twin) that is related to the response (the height difference). Making the example more realistic by allowing for variation of male and female heights would not change the conclusion, because the expected bias depends only on the means and not on the variability.

The analogy with a cross-over rainfall enhancement experiment is immediate. Instead of observations on pairs of twins, think of observations on a pair of targets; instead of a twin selected for treatment, think of a target selected for seeding; instead of gender, think of cloudiness, with M indicating presence of suitable clouds and F their absence; and think of 175 and 165 as amounts of precipitation on the two targets, the former occurring on “ M ” days (with suitable clouds on target), the latter on “ F ” days (without such clouds on target). Design D1 then enrolls all days, design D3 enrolls only days with suitable clouds *on both targets*, and design D2 enrolls only days with suitable clouds *on the selected target* (which is precisely what experimenters are often clamoring to do). Design D2 is seen to yield a biased cloud seeding experiment.

A separate issue is that unbiased design D3 will usually give greater precision than will unbiased design D1, because D3’s units are more homogeneous. Conclusions from D3 will, however, be of less general applicability than conclusions from a D1 experiment.

These warning examples show the danger of bias resulting from the inclusion of units being dependent on the allocation of treatment. To avoid such biases, definitions of units and their enrollment must be independent of the treatment assigned. Thus, in the Whitetop experiment, the envelope containing the assignment “was opened *only after* we had completed all actions and decisions relating to designation of an experimental day” (Braham 1979, p. 60). A possible alternative way of excluding unsuitable units is to sift them out *ex post*

facto by a criterion that is independent of the treatment, as was attempted in the analyses of the Israel and Puglia experiments, which excluded days that had had no precipitation at stations in a buffer zone.

If the treatment did affect selection, even indirectly, it might entail bias. The reason one may still, despite these potential sources of bias, sometimes use units whose definition depends on responses is because they may be physically meaningful. An example is defining a cloud over its lifetime, even though the persistence of the cloud might be affected by seeding.

On the other hand, some criteria of exclusion are so clearly independent of treatment that the excluded units can be omitted from the analysis without risk of bias. Thus, patients lost because of accidents or other events unrelated to the study can surely be excluded (Senn 1997, section 11.2.3), as can weather units during which the seeding equipment was not available (Gabriel 1967). A less obvious criterion of exclusion is one depending on concomitant measures such as the absence of precipitation in an area adjacent to the targets but never targeted for seeding. Its justification in the rain enhancement experiments in Israel and in Puglia was that the adjacent area "was chosen for this classification because it can hardly ever be affected by seeding operations . . . and hence . . . is most unlikely to introduce bias" (Gabriel 1967, p. 97; List et al. 1999).

e. Treatments and comparison

Treatments in clinical trials are compared either with other treatments (innovative to standard) or to lack of treatment. The latter comparisons run the risk of being biased by what is called the placebo effect and consists of the well-known optimistic effect of the very existence of a trial on both patients and medical personnel. Trials therefore usually assign placebos (inert substances) to the patients not receiving the treatment. Moreover, when possible, both patients and attending physicians, as well as technicians who observe and evaluate patient responses, are blinded to whether the treatment assignment to a particular patient has been active or placebo. Experimenters go to great lengths to ensure patient blindness to treatment, even to the extent of boring holes in the crania of all patients in a Parkinson's disease study, so patients did not know whether they had received brain surgery (New York Times 1999). The other aspect of blinding is easier, that of having patients' lab tests, X-ray scans, and so on assessed in ignorance of the identity and treatment of the patients. Elaborate administrative arrangements ensure this blindness and identify patients' treatment only at the stage of the final analysis. Sometimes this is not possible, as in a study of the effect on blood clotting of using an anticoagulant versus using an external massaging device; and there then is an inevitable problem of possible bias (Francis et al. 1992).

Almost all weather experiments have compared seed-

ing, usually with silver iodide (AgI), with lack of seeding, though one could compare different seeding techniques, such as glaciogenic versus hygroscopic, or higher versus lower dosages (Gabriel and Changnon 1982), or seeding on one target versus seeding on another. Comparisons of seeding with no seeding might be subject to placebo effects of the turbulence introduced by penetration of aircraft and release of materials, as well as by the inevitable bias of the people who record and review observations. To avoid these biases, some experiments have used flares with sand on occasions assigned not to be seeded, and thereby blinded the observers on the aircraft to the distinction between seeding and not seeding (Woodley et al. 1982; Changnon et al. 1995). For the same reason, interpreters of radar scans were blinded to the occasions AgI was used; it was not enough to remove the dates from the radar scan records, which would have paralleled the way X-rays are assessed in clinical trials, because meteorologists might well have been able to identify some of the dates from the radar scan. Execution of such a blinded protocol is administratively difficult, especially in cloud seeding. It is, for example, not easy to assign AgI or sand (placebo) flares to a particular seeding run without the pilot and seeding officer being aware of what is loaded on the aircraft.

f. Randomization

To protect against any unforeseen sources of bias, the actual assignment of treatment to each unit must be done at random, that is, according to a probability process. One way to do this is by randomizing the assignment after the unit has been designated, essentially by opening an envelope containing the random assignment. Another way is to create a sequence of randomizations ahead of time, sequester it so as not to influence the selection of units, and only make it available, one assignment at a time, after each new unit has been designated. In clinical trials, patients who arrive and present with certain conditions are first definitively enrolled, and only then is their assignment randomized by calling in at a preprogrammed computer. One meteorological example is the Swiss Grossversuch III in which forecasts of potential hail occasions were used to enroll days as experimental units, and then the top envelope was chosen from a stack that had been prepared randomly with "ja" or "nein" cards, and seeding was carried out if the envelope contained a ja card. In the Israeli experiments, on the other hand, the list of random assignments could be published ahead of time because no days of the rainy season were to be initially excluded, and seeding allocation could not affect the subsequent exclusion of days that had no rain in an always unseeded zone.

Randomization not only protects against bias but also provides a probability model that allows the assessment of the chance of various outcomes under particular assumptions regarding treatment effects. It allows the cal-

ulation of significance levels under the null hypothesis of no effect, as well as the estimation of power under hypotheses of effects of given magnitudes. In this respect there is no difference between clinical trials and weather experiments, because the probabilities depend on the stochastic manner in which randomization is carried out, not on the medical or physical world to which they are applied.

g. More on the need for “black box” experimentation

Research workers are continually learning more and want to incorporate their new findings into the ongoing research, adapting treatments and observations as they go. Can that type of adaptive experimentation be consistent with the methodology of evaluating the results of designed experiments?

If the change in definition is driven by the observations in the earlier part of the experiment, applying it to the earlier data obviously biases the results. Some such issues arose during the execution of the Israel I experiment, as follows. A change in seeding line from that originally planned was requested by the pilots who wanted to stay within sight of the coastline. This change was judged unlikely to produce an important change in treatment effect and was therefore incorporated in the experiment, as were changes in the hours of starting and ending each experimental day. On the other hand, a proposed change of target boundary lines to exclude the coast and the Jordan Valley was resisted because it was driven by observations on the earlier part of the experiment, and thus might have introduced a bias (Gabriel 1967).

Changes that are driven by observations during earlier parts of a designed experiment must be avoided because they can create bias. If such changes are crucial, the original experiment must be terminated and a new experiment designed. If the new technique is really superior to the older one, why should its effects be diluted by analyzing them with the results of the older technique? At the very least, the experiment should be analyzed within separate strata for the different techniques. Whether or not the parts can be analyzed jointly, as strata of one study, is moot. Even if they can be analyzed jointly, this method requires the experiment to be lengthened. An analysis after a single change increases the number of tests from one to three: one test of the pre-change effects, a second of the postchange effects, and a third test of the difference between them. To obtain reasonable power for each test, the experiment would need to be prolonged very considerably. Allowing more than one change augments the number of tests yet more and requires impossibly long experiments.

Returning to medical examples, it is useful to quote an experience at the University of Rochester School of Medicine.

I just finished participating in a clinical trial in which,

half-way through (but without any knowledge from any analysis of the available data) a new version of the treatment under test became available. Should we scrap the “half trial” and start a new one? The decision was made, based on the assumption that the effectiveness of the new device was identical to that of the old version—an assumption the [physicians] felt strongly about—to continue but to introduce a new stratum, consisting of all patients recruited when the new device was available (treatment and control). So the experiment was lengthened to yield extra power and a stratified analysis was done. Of course, we also tested the secondary hypothesis of whether the two versions differed in their effectiveness, but that was largely “for show” since little power was available. . . . But if the change of the protocol had been based on *results* from analyzing the early part of the trial, I wouldn’t have known how to proceed.

(W. J. Hall 1996, personal communication)

Adaptive experimentation must be separated from confirmatory analysis. Research workers who feel they must continually learn and adapt their techniques cannot ask for statistical confirmation. Decisions on treatment and/or policy may impose the choice between betting on the research workers’ most recent ideas or insisting on designed experiments with adherence to rigorously defined protocol. The tension between subject-area scientists and statisticians is that between inspiration and skepticism, and it is a difficult task to maneuver between them.

4. Responses and covariates

a. Definition of the response

In clinical trials it is often difficult to define a response that is both meaningful and readily measurable. Some responses, such as death or survival, are fairly straightforward. But survival for how long? Cancer treatments are often assessed in terms of 5-yr survival, but for slow-developing cancers, 10-yr survival may be more relevant. An alternative is to define the response in terms of length of survival. That definition, however, depends not only on the disease and treatment that are the subject of the study but on other causes of death as well, so the response may be defined in terms of length of remission. That definition raises another problem: What should be done with people who die from other causes while in remission? How long would such a person have been in remission if they had not died from the other cause? Such problems have generated a branch of statistics called survival analysis (Cox and Oakes 1984), which applies not only to data of human survival but also to data of product endurance, and might be useful for the analysis of lifetimes of meteorological phenomena such as storms, clouds, or echoes.

Other clinical trials focus on recovery, on cessation of symptoms, and so on and are fraught with problems of definition. For example, in clinical trials of analgesics

the response must be pain, something that is very difficult to measure objectively. An attempt to do so for postsurgical patients is to compare the patients' self-evaluations some time after receiving the analgesic (or placebo) with their self-evaluations before receiving it (Cox et al. 1980). There is some analogy here with the use of damage claims in assessing hail prevention trials.

In weather experiments, the choice of response is fairly straightforward when experimental units are defined by an area and a period, largely because the rationale for those units is that precipitation measurements are available for them (such as networks of rain gauges and/or hail pads). For units such as clouds or storms, on the other hand, the response is usually a complex evaluation of radar records, as in Illinois (Changnon et al. 1995) and Thailand (Silverman et al. 1999; Woodley et al. 1999), and attempts are even being made to trace the hail and rainfall from a storm by tracking its path with the help of sophisticated computer interpretation of radar records (Berthoumieu et al. 1999).

b. Surrogate responses

The clinical outcome of a trial on survival, or on the onset of acquired immunodeficiency syndrome (AIDS), may take a long time and great cost to observe. To reduce duration and cost, therefore, trials often use surrogate endpoints such as the reduction of cholesterol levels or CD4 cell counts, respectively. This approach, however, runs the risk of being misleading if the effect of treatment on the surrogate is not the same as on the clinical outcome. Thus, trials on heart disease have shown that treatments did reduce cholesterol levels but did not reduce mortality, and trials on human immunodeficiency virus therapy have shown effective reduction of CD4 counts without a corresponding improvement of survival (Fleming and DeMets 1996).

There are obvious analogies with weather modification. Possible increases in precipitation do not translate simply into economic benefits, and there is much uncertainty in translating augmented radar reflectivities into increased precipitation on the ground, and a fortiori into economic benefits. [Studies of some of these relationships have been attempted for some time (e.g., Changnon et al. 1977).]

Surrogate responses may be relevant to understanding the science underlying the treatment effects such as in phase-II screening trials (Fleming and DeMets 1996) and in basic meteorological research. But one needs to be careful in extrapolating findings on surrogates to the really intended effects.

The concept of surrogation is carried even further when the experimental units are laboratory animals rather than human subjects. The ethical problems of carrying out experiments on people make it important to obtain as much information as possible from such uses of animal surrogates, albeit finding effects on animals

is not conclusive about the existence of similar effects on humans.

Parallels in weather experimentation are simple experiments in clouds, such as that of Langmuir and Schaefer, as well as some laboratory and computer model experiments. Unfortunately, no other unequivocal experiments with clouds have been possible, and the reduction of scale from the cloud to the laboratory is too large to allow ready transference of problems and results. There is more hope from computer models that reproduce properties of clouds as closely as is possible within the realm of what is understood of atmospheric physics (Orville 1996). Given the similarity of the models to real clouds, it becomes of interest to see the effect of simulated seeding on the models (Yin et al. 1999) and to study such things as the differential reaction to different methods of seeding (Levin et al. 1999). It is well, however, to keep in mind that the behavior of these models may be no more conclusive of the behavior of clouds than the reaction of rats is conclusive of the reaction of humans.

c. Multiplicity of responses

There usually are several responses that may be relevant to any one study. An experiment designed for hail reduction might also be analyzed as though it had been a precipitation enhancement experiment, as was done with Grossversuch III (Schmid 1967). In the analysis of precipitation, the analysis may focus on initiation of precipitation as well as on amount of precipitation, and on the growth or lifetime of a cloud. Radar echo may be considered a relevant measure, as is precipitation on the ground. The response may be measured immediately after seeding, after a delay, or after each of several delays. Precipitation on the ground may be measured by several networks of rain gauges.

Investigators who test effects on many responses, say one hundred, can expect, when the treatment has no real effect at all, to find that some five of the responses are significant at 5%. They are likely to be tempted to publish those responses, conveniently forgetting the non-significant 95 or so. This *response multiplicity bias* can be avoided only if a single response, or a specified few responses, are chosen ahead of time or at least independently of the data. It is not easy to persuade researchers to do so, because there are so many possible effects they would like to study. One way to maintain the integrity of the analysis is to adjust the level of significance to take account of the number of analyses planned beforehand (Multicenter Diltiazem Post-Infarction Trial Research Group 1988). Awareness of how much this adjustment reduces the level may also dampen the investigators' enthusiasm for testing too many hypotheses.

There is great temptation to focus the analysis on a response that is found to be "promising" during the experiment, but if the response is not chosen at the

design stage, one may easily bias the analysis by letting the data guide its choice, that is, by effectively choosing to analyze the “most promising” of the many available response measures. Experimenters often find it difficult to see how this potential bias applies to their study, but they should think of the analogy of how frequently some “coincidence” seems to occur in life, especially when one ignores the much greater frequency of a particular coincidence not occurring.

d. *Covariates*

Covariates are measures available for the experimental units that are (i) unrelated to treatment allocation, and the most obvious instances are measurements taken prior to allocation, such as baseline medical observations or weather records taken prior to cloud seeding. Covariates can improve the efficiency of an experiment if they are (ii) correlated with the responses, such as baseline measures with similar response measures after treatment and preseeding radar reflectivities with postseeding reflectivities. They improve efficiency because they allow treatment comparisons to be made between experimental units that are similar to one another, that is, between units that have similar covariate values but different treatments. They do so at the design stage by allowing the units to be grouped into relatively homogeneous subgroups (blocks, strata), and at the analysis stage by adjusting the responses for the covariate values (analysis of covariance, ratio statistics). Either way, they serve to reduce the noise against which the treatment signal is to be evaluated.

Clinical trials have obvious covariates in sex and age at enrollment, because these traits are related to most responses and are unaffected by treatments. Thus, (ii) men respond to pain differently from women, and the young survive longer than the old, but (i) sex and age at enrollment are unaffected by medical treatments. It has been suggested, however, that covariates are of doubtful usefulness in adjusting statistical analyses of clinical trials because the enrollment is usually limited to a very homogeneous group of patients in which correlations with covariates are very low (Beach and Meier 1989). The situation may be similar in weather experiments with cloud units, or storm tracks, because it is difficult to find covariates that are sufficiently highly correlated with their response and yet unaffected by the treatment (see, e.g., Changnon et al. 1993, chapter 6; Bradley et al. 1980). An exception is the reanalysis of the South African glaciogenic flare seeding experiment (Silverman 1999) in which the postseeding radar echoes (responses) were usefully regressed on preseeding echoes (covariates).

Covariates have been very useful in adjusting analyses of weather experiments whose response was target area precipitation in a fixed time period. The best covariates have been precipitation in nearby areas, preferably upwind. In choosing them, one must compromise

between closeness to the target, which yields (ii) high correlation, and sufficient separation to ensure (i) they cannot be affected by seeding aimed at the target. Also, there is hope for the development of better covariates calculated by introducing prior data into physics models that might be honed to give close predictions to subsequent responses.

Some variables need elaborate and lengthy working up of measurements and therefore become available only after the treatment has been applied. Such variables can serve as covariates so long as they are calculated independently of the treatment allocations, possibly by ensuring that the persons calculating them are blind to the allocations. It is, however, not always obvious whether a potential covariate might be affected by treatment. In medicine there is the temptation to use the amount of medication, or the length of time the medication was administered; in weather modification there is a similar temptation to use the duration of seeding. The only way these can be legitimate covariates is if the decision on amount or duration of treatment is made blindly to the treatment assigned. In medicine this situation happens when the doctor regulates the amount of medication without knowing whether the patient receives treatment or placebo, and in cloud seeding it happens when the decision to terminate seeding a unit is made in ignorance of whether it is being seeded with AgI or with sand. A more subtle concern arises if the response depends on a storm's length, but seeding might prolong a storm, so the length does not qualify as a covariate. Again, “time of initiation” of seeding may be a poor covariate because it is not well defined on occasions assigned to not seeding, or assigned to be seeded but with conditions changing by the disappearance of clouds that were assigned to be seeded.

5. The statistical design of an experiment

The statistical design of an experiment defines the allocation of the treatments to the units in terms of a model that will guide the analysis.

a. *Blocking and stratification*

The units might be grouped into relatively homogeneous blocks or strata, so that much of their variability is between these groups and little remains within the groups. If the treatments can be expected to have the same effects in all groups, that is, if group differences and treatment effects are *additive*, experimental results can be analyzed by making treatment comparisons within each block or stratum and averaging these comparisons for overall estimates of the common effects. Grouping might be carried out prior to treatment, in which case randomization should be made within the groups (see section 5f, below), or it might be implemented only at the stage of analysis. Either way, it is a valuable feature of an experimental design provided

the blocks or strata are relatively homogeneous so that within-group comparisons are subject to less variability than overall comparisons would have been in an ungrouped design.

Blocking has been effective in medical research, especially in laboratory experiments in which there are day-to-day, week-to-week, or interlaboratory differences in calibration of instruments. Blocking by day, week, or laboratory then reduces variability and yields a valid design provided that treatment effects are constant over days, weeks, or laboratories. Thus, a surgical intervention may be expected to be equally effective in all hospitals and from year to year, so a study could be blocked by hospitals and/or by years. Blocking has yet to be found useful in weather modification because it is difficult to create homogeneous groups of meteorological units and even more difficult to assume cloud seeding to have the same effect in all of them. "I would definitely question anyone's ability to collect experimental units into relatively homogeneous blocking when it comes to cloud classification, related cloud seeding applications, and resultant effects," (T. Henderson 1996, personal communication).

b. Treatment-group interaction

If different groups of units are affected differently by the treatments, this is referred to as *treatment-group interaction*, or *nonadditivity*. It then makes no sense to pool the results from different groups, and the experiment really consists of several subexperiments. It therefore is necessary, whenever differential treatment effects are suspected, to subdivide the data so that one may expect a reasonably constant effect within each division.

An example of treatment-group interaction in medical research is the difference found between the Diltiazem-related reduction in cardiac events among patients without pulmonary congestion and the corresponding increase in such events among those with pulmonary congestion (Multicenter Diltiazem Post-Infarction Trial Research Group 1988). In weather modification, the data of several experiments were broken down by covariates such as cloud-top temperature and wind direction (Gabriel and Baras 1970) and by such synoptic criteria as cold fronts and air masses (Changnon et al. 1993, chapter 4), because it was suspected that the effect of cloud seeding might interact with these groupings. In the latter Israeli experiments, data were analyzed separately for days with and without desert dust, because it was surmised that the effect of seeding changed with the presence of such dust (Rosenfeld and Nirel 1996).

c. Cross-over experiments

A design that has been much used in medical research is that of using units repeatedly, crossing over from a period with one treatment to a period with another. The potential saving in units required for an experiment

makes this design very attractive, but concerns with carry-over of effects limit its usefulness (Hills and Armitage 1979; Senn 1993, 1997, chapter 17). This kind of design has been used in weather modification by means of alternation of target areas, each one being seeded on some occasions and compared with the other, which then serves as a control (Moran 1959; Gabriel 1967). Recent discussions of the cloud seeding experiments in Israel and Puglia have raised the possibility of differential effects on the two targets (Gabriel and Rosenfeld 1990). If that were so, it would vitiate the rationale of this design, which is tailored to estimate a common effect.

d. Multicenter trials

Clinical trials are often carried out simultaneously in several medical centers so as to accumulate enough units within a reasonable length of time (Senn 1997, chapter 14). The advantages of such pooling of resources are obvious, but it is justified only if treatment-center interaction is negligible. Unfortunately, the power of detecting such interaction from the data is usually very low, so the validity of inferences from multicenter trials often cannot be verified.

The idea of multilocation weather experimentation has been raised by W. L. Woodley (1999, personal communication) with a view to rapid accumulation of an adequate amount of data on a cloud seeding technique. The possibility of doing so will depend on how reasonable it is to assume that the technique will have similar effects at different locations. The meteorological experience with cross-over designs suggests, however, that effects may be suspected to vary even between nearby areas, and a fortiori between distant locations.

In either kind of experimentation, the rationale of multicenter studies hinges on the existence of communality of effects in the different centers; otherwise, there is no point in running the several centers' trials together. As of now, there appears to be more justification to assume common treatment effects on patients at different centers than common cloud seeding effects on clouds at different locations.

e. Restricted randomization

Given blocking or stratification, the treatments must be randomly assigned to the units *within* each separate block or stratum rather than to all experimental units at once. With small blocks, however, this requirement may result in unbalanced assignments, with most or all of a block's units receiving the same treatment. If, for example, the blocks had two units each, simple randomization of treatments A and B would assign AA or BB to about half of the blocks, and these would provide no information on treatment effects. This result might be avoided by allowing only balanced assignments; that is, each block of two units would be randomly assigned

either AB or BA (AA and BB not being assigned at all). That, however, could violate blindness because once the experimenters knew that the first unit of a block had received A they could conclude ahead of time that the second unit would receive B, and vice versa. Without blindness there might be selection bias, as in the case of a very sick patient not being enrolled because it was known placebo would be assigned, or a day with a low chance of rain not being included because it was known seeding would be assigned.

One way to reduce this kind of imbalance in an experiment with small blocks is to use restricted randomization. An example is the current hail prevention experiment at Agen, France (Berthoumieu et al. 1999) in which allocation of treatment A or B is randomized as follows. When the n th unit comes up for randomization, if the difference between numbers of earlier allocations to A and to B was less than 2, randomize with probability 1/2 for A; otherwise, if there were more allocations to A than to B, randomize with probability 1/3 for A; otherwise, if there were fewer allocations to A than to B, randomize with probability 2/3 for A.

6. Statistical analysis

The method of analysis must be part of the initial design; that is a necessary condition for the validity of statistical inferences. If the method of analysis were chosen so as to highlight features first observed during the experiment, the resulting calculations of tests and confidence intervals would be devoid of probabilistic meaning.

a. Rerandomization

Given the design and sufficient knowledge of the distribution of the response and how it could be changed by the treatments, methods of mathematical statistics could be used to derive statistical tests of maximum power. Thus, Neyman advocated a class of procedures, referred to as $C(\alpha)$ tests, that would be optimal if, among other things, treatment increased the response variable proportionately to its untreated value; no one, however, has ever produced any evidence that cloud seeding effects are of that type. Indeed, most weather modification experiments are designed and analyzed with little reliable knowledge of response distributions and treatment effects, so that the methods of optimizing power are of doubtful relevance. To quote Kempthorne (1980), "I do not attach high value, vis-à-vis weather modification, to work on the mathematics of testing hypotheses on various mathematical distributions, though again this may be useful with regard to suggesting test statistics that one will examine via the experiment randomization distribution."

It is now generally accepted in meteorology, especially since the forceful advocacy of John Tukey (Tukey et al. 1978), that valid statistical inferences necessitate

rerandomization analysis or large sample approximations of such analyses (Gabriel and Feder 1969; Tukey et al. 1978; Mielke 1985; Gabriel 1999). [For discussions of particular statistical techniques the reader is referred to Dennis (1980); Wegman and DePriest (1980); Mielke (1985); Gabriel (1999); and numbers 10 and 11 of the 1979 volume **A8** of *Commun. Stat. Theory Methods*.]

Medical studies make greater use of distributional assumptions and of statistics that have optimal properties under these assumptions. Are distributions of medical variables so much better known, or less given to extremes than those of the weather; is the kind of effect of treatment so much better understood; or are there also doubts in medical research about the justification of using statistical techniques that are optimal only under special conditions?

b. Statistics that exploit features of the design

In the absence of trustworthy assumptions about distributions and effects in weather experiments, it is reasonable to use statistics that are intuitively appealing and exploit the variance reduction features of the experimental design. Such are analyses of covariance and ratios of responses to covariates, because these techniques allow the adjustment of responses to covariates (Moran 1959; Gabriel and Feder 1969; Gabriel 1999). Regression methods, including generalized regressions with various links, are similarly used in medical statistics to take out the influence of covariates such as age or baseline measures, and many medical analyses use nonparametric or semiparametric modeling that utilizes only limited assumptions.

c. Length of an experiment

Clinical trials are designed to run long enough to give sufficient chance (power) of discovering a real treatment effect if there is one. The required power depends on the needs and purposes of the experiment. This requirement translates into the number of suitable patients required and the time it will take to enroll, treat, and follow them up. In weather experiments with a cross-over design of daily units, experience shows that about 5 yr are needed for 90%–95% power of finding a 5% significant result if the true seeding effect is a 15% increase of rainfall (Gabriel 1999). Other experimental designs, such as those with a single target and control, require considerably more time, as does the detection of effects of less than 15%. Shorter experiments might be possible if covariates were available that would seriously reduce the variability of the adjusted responses.

d. Early termination of an experiment

An important issue is the temptation to peek at the intermediate results of experiments and declare a sig-

nificant outcome as soon as the calculations indicate a P value of, say, 5% or less. The probability that such calculations will show a P value less than 0.05 *sometime* before the experiment has run its planned course is considerably greater than 5%, so this peeking strategy can greatly bias the conclusions and has been generally discouraged by statisticians.

In clinical trials, the temptation to peek is motivated not only by curiosity and the wish to reduce the cost of experimentation, but also by the ethical obligation to announce the effectiveness of a new treatment as soon as possible, and, on the other hand, to abort trials when concerns arise about the safety of the treatments. In the well-known trials of azidothymidine (AZT) treatment of AIDS patients some years ago, intermediate results suggested greater survival for those receiving the drug. This finding was announced and the trials had to be terminated because once a treatment has been considered effective it becomes unethical to continue an experiment that administers placebo to some patients. There had been much soul searching about this decision among the investigators, who had to weigh the advantage of administering AZT to then current patients *if it really was effective* against the disadvantage for future patients of having less reliable information about AZT from a sample of reduced size. In retrospect, the decision about AZT appears to have been wrong, but the judgment call, at that time, may well have been just.

Statisticians have now devised more flexible rules that permit some peeking at intermediate points without biasing the decision. As an illustration, consider annual peeks at a 3-yr experiment. An overall 5% significance level could be ensured by the following O'Brien-Fleming (1979) rule: If the first year's results exceeded the 0.06% significance level the trials would be terminated and significance declared; otherwise, if the second year's results exceeded the 1.51% significance level the trials would be terminated and significance declared; otherwise the trials would be carried on to their 3-yr conclusion and significance declared if the results exceeded the 4.71% level. This rule is attractive because it does allow the trials to be terminated after one year if the first year's results are very strongly indicative of an effect; it further allows the trials to be terminated after the second year if the results up to that time are strongly indicative; yet its chance of attaining significance at the end of the three years is almost as high as that of an ordinary no-peek 5% rule. [This and other rules are described by Piantadosi (1997), chapter 10.]

A similar strategy might be considered for some weather modification experiments, though the importance of early decision may not be as great there as in medicine. Unfortunately, considerations of early termination of weather experiments have been driven by apparent lack of success rather than overwhelming early success. Early termination of an experiment is at times indicated when intermediate results, such as those in Puglia after some 85% of the planned units had been

observed, show that it is highly unlikely that significance could be achieved by carrying out the experiment to its intended end (List et al. 1999). Similar problems in medical research have been discussed under the heading of *stochastic curtailment* (see, e.g., Davis and Hardy 1990).

A difficult call is how to evaluate the results of experiments that have been terminated prematurely for reasons that did not depend on its observations, such as exhaustion of funds. It is appealing to evaluate them as though they had been planned for the length they actually ran, but that presumes one really is sure the results had no effect on the termination. Can one really ever be sure that if the early results had been more promising the policy makers would not have been prevailed on to support additional years of experimentation? This area is obviously a gray one.

e. *Exploratory analyses*

Experiments yield large collections of data that may contain clues about detailed treatment effects. Sifting through such data can yield ideas and suggestions, but also entails the danger of invalid inference when a serendipitous observation is treated as though it confirmed an anticipated phenomenon. Noticing one drug out of a hundred having an apparent effect does not mean the same as finding the same effect on the one particular drug on which it was expected. Locating increased precipitation at any one out of a hundred locations in the target does not have the same evidential value as finding it at the one location that had a priori been predicted to be the most likely place to be affected by seeding.

Dredging data from adaptive or designed experiments for clues and new ideas is referred to as *exploratory data analysis*, as distinct from the *confirmatory analysis* that tests an a priori hypothesis by methods laid down in the initial design. The latter is subject to rigorous formulation by mathematical statistics that permit valid probabilistic statements on significance tests and confidence intervals. The former is an a posteriori examination of the data that may provide ideas about the phenomena under study. Applying significance and confidence statements to exploratory analyses is a common and insidious pitfall and easily leads to the propagation of type-I errors, that is, to false declarations of significance when there really is no effect.

The data accumulated in clinical trials include demographic information about the patients as well as numerous secondary outcome variables and seemingly innumerable laboratory data, all of which *may* have relevance to the trials' objectives. They may provide invaluable help in suggesting unanticipated associations of treatments with a variety of responses (Viagra was studied for its effect on the heart when its effect on another organ was noticed) and possibly also interactions with various groupings. Such suggestions, how-

ever, need to be tested independently by experiments designed for that response.

Weather experiments similarly accumulate details of synoptic conditions, measurements aloft and on the ground, radar scans, seeding logs, and so on, all of which are possibly relevant to the effectiveness of cloud seeding. These data can suggest new ideas, as in the example of cloud-top temperature windows in which glaciogenic seeding appeared to be effective (Gabriel and Baras 1970); they can also modify old ideas, as in explaining apparent negative effects of seeding by the presence of desert dust (Rosenfeld and Nirel 1996). They may suggest that an apparently unsuccessful experiment contained a subset of occasions with highly successful seeding, and they may equally suggest that some claimed effects of seeding might be purely chance results of shifting weather systems (Rangno and Hobbs 1993, 1995; see, however, Gabriel 1995; Mielke 1995). All these, however, have to be viewed as merely *suggestive*, and attachment of calculations of “significance levels” does not raise them to the level of confirmatory evidence. That level can be obtained only by designing and carrying out a new study.

There is a gray area of cross-confirmation by statistics and subject area, when data of doubtful statistical significance are combined with an uncertain rationale from the subject science. Critique of the statistics is muted by the supposed strength of the science, and the doubts of the subject-area scientists countered by the supposed strength of the statistics. There are no hard and fast rules to resolve such situations, and impassioned argumentation is no help. Nor can there be a recipe on how to deal with a statistical finding that is not supported by any reasonable medicine or physics, as the case may be. When such situations are met, doubts must prevail until they are resolved by further experimentation.

Another gray area is the confirmation of the exploratory findings in one experiment by those of another experiment. Another experiment may, indeed, provide some confirmation if the conditions and definitions are the same, but how does one know whether they are? Confirmation by especially designed experiments is much more convincing.

Given the concerns that have been voiced about merging strata in one experiment when it is doubtful that the effects are the same in all strata, it is difficult to be comfortable with meta analyses that implicitly assume effects to be equal over space, seasons, techniques, agents, and so on. Ambitious but naive techniques of pooling results should not replace intelligent comparisons of separate results. [For critical discussions in the medical context see Oakes (1993) and Senn (1997, chapter 16).]

The wide-ranging possibilities of exploratory data analysis can be a fruitful source of ideas and hypotheses, but precisely because they are so wide ranging, they cannot either *confirm* any hypothesis or *reject* it. The fundamental idea of scientific inference is that prefor-

mulated hypotheses must be tested on independent data, and that idea does not allow a posteriori custom tailoring of hypotheses to fit a given dataset.

7. Analysis of operational data: Historical comparisons

Much information is gleaned from the analysis of data collected in the course of medical practice and in the course of commercial seeding operations. The ever-present question is how much credence can be given to inferences from such “historical” data. No one would deny the validity of inferring the reduction of mortality as a result of the spread of better health practice, even though the data are nonexperimental, but there is great doubt about the causal interpretation of a recently observed correlation between the use of Aspartame and the frequency of brain tumors. Then there is global warming: Is it a continuing trend, and is it caused by human activities? It is difficult to be sure of the causal interpretation of historical connections, but it is impossible to ignore such evidence, especially when rigorous experiments are unavailable or impossible. The most extensive data on cloud seeding operations are probably those of 250-odd project seasons of a large number of nonrandomized projects carried out by Atmospherics, Inc.

Every project has been analyzed by using some type of statistical methodology. Many of these evaluations have been conducted by federal agencies, universities, state and county groups, program sponsors, and a few private statisticians. Most of these analyses “suggest” positive results. . . . Does any of it contain anything of real value? . . . I lean toward “yes.” How confident am I in this belief? Probably about as confident as *anything* I happen to believe, given the realization that very few decisions I make on any subject are based on a randomized experiment. (T. Henderson 1996, private communication)

An example of the problems that recur with nonexperimental data is that of the treatment of prostate cancers—a frequent question is whether radiation is better than surgery, or vice versa. Each of them is highly recommended by its practitioners (Brendler and Walsh 1992; Epstein and Hanks 1992). The statistics of extensive follow-ups of patients who received one or the other treatment show the general survival rate to be somewhat greater after surgery. But the decisions as to which patients get surgery and which receive radiation are made by physicians, whose usual practice is to recommend the more radical treatment, that is, surgery, for younger and stronger patients, and to recommend radiation for older and weaker patients. No wonder that the statistics show better results for surgery patients!

Bias might be reduced by basing the statistical adjustments on independent assessments of stage and severity, that is, by assessments not carried out by the patients’ physicians. It may, however, not be possible

to do this routinely in a nontotalitarian society. It therefore remains extremely difficult to make a definitive comparison of surgery and radiation, even though huge amounts of prostate cancer data have been accumulated by practitioners of each therapy. In the absence of randomized assignment of treatment, it may never be possible to be certain which is better.

The parallels with weather modification research are obvious. The assessment of the effect of cloud seeding during nonrandomized operations is based on comparisons with historical data and/or data from other areas. How can one ever know whether the difference between the operational precipitation and that in other places and other times is due to seeding? An obvious source of bias is that seeding operations usually occur after a drought and address the most affected area. It is only to be expected that the seeding period has relatively high precipitation in comparison with the preceding period. Cases of similar "regression to the mean" are also well known in medical studies (Senn 1997, section 3.5).

Data from cloud seeding operations must be used very circumspectly, and can never have the same evidential weight as do data from randomized experiments. This fact has been well known and documented since the early days of weather modification in the 1960s (Neyman 1977), and it is well understood that operators cannot produce conclusive evidence of the effect of their operations. The dilemma for operators is that their funding often depends on seeding on every available occasion, so they can never leave an occasion unseeded; as a result, they cannot produce reliable evidence on whether their seeding is effective. A possible "compromise" is to design experiments that can be piggy-backed on operations by randomization of differential dosages and/or agents (Gabriel and Changnon 1982).

8. Practical relevance of experimental findings

a. Technological innovation and the time lag to experimental confirmation

Returning to the example of cancer therapies, it had been of interest to consider another treatment option, that of brachytherapy, in which minute pellets of irradiated gold, palladium, or other metals are implanted into the organ about the location of the cancer. Radiation then emanates from inside the organ and is focused more directly on the cancer, with fewer harmful side effects on surrounding organs. It continues for a few months—the half-life is some 60 days—by the end of which it is hoped that the cancer is destroyed.

In the early 1990s, however, the statistics available on brachytherapy showed a very low 10-yr remission rate. Such statistics could be calculated only some 10 years or more after treatment and therefore at that time were available only for patients who had been treated in the 1970s and early 1980s, when implantation required a major operation and placement of pellets was

very inaccurate. By the early 1990s, however, irradiated pellets could be placed without surgery, using real-time imaging of the organ, and their implantation at the cancerous site was much more precise (Nori 1993). Were the statistics available at that time relevant to this newer method of implantation? Probably not, and they were therefore not relevant to the choice of this treatment at that time.

Similar concerns arise in the study of cloud seeding technologies. Are the results of studies of glaciogenic seeding by AgI relevant to the present possibilities of hygroscopic seeding? The intended physical effect is very different, so why should the observed effects on the ground be the same? *Are statistics about applications of technology always irrelevant, because they relate to past performance rather than to present capabilities?*

b. Scientific confirmation and practical decisions

Serious doubts persist regarding the effectiveness of cloud seeding for augmenting precipitation, especially with glaciogenic material, and yet some seeding operations may be economically justified. Why not risk a relatively small outlay on seeding clouds if it is considered at all plausible that it will result in increased precipitation or reduced hail, both of which are of considerable economic value. After all, other business decisions are also made without "proof" that they will turn out to be profitable; economic choices generally are decisions under uncertainty.

The accepted relation between proof and application is very different in medicine. Administration of a treatment is contingent upon rigorous demonstration of its effectiveness. Drugs cannot be marketed unless they have been found to be significantly effective in clinical trials. Patients are permitted to choose a treatment not if they are convinced of its efficacy but only if that efficacy has been confirmed by rigorous experimentation.

The standards for scientific demonstration are very distinct from those for taking action. In meteorology it is accepted that the latter are determined by market forces: Seeding operations can be conducted absent scientific proof. In medicine, by contrast, it is accepted that drugs and treatments be made available only if there has been scientific confirmation of their effectiveness. The difference between the public attitudes toward science and policy in these two areas shows that this is not a matter of scientific method or principle but a reflection of what aspect of life one wants protected by governmental and judicial regulation. Evidently, people feel less need to be protected from possibly futile interference with the weather than from possibly unavailing medical treatment.

Acknowledgments. This paper is based on a talk presented at the International Workshop on Regional Pre-

precipitation Enhancement held at Bari, Italy, on 11–15 November 1996. It owes much to numerous discussions with Bernie Silverman through many years and to encouragingly critical comments by Ted Henderson, Jack Hall, Mike McDermott, David Oakes, Rich Raubertas, and Stephen Senn.

REFERENCES

- Beach, M. L., and P. Meier, 1989: Choosing covariates in the analysis of clinical trials. *Controlled Clin. Trials*, **10**, 161S–175S.
- Berthoumieu, J.-F., A. Loretz, A. Carlier, F. Abdellani, E. Lambert, J.-R. Mathieu, and K. R. Gabriel, 1999: Cloud base hygroscopic seeding to reduce hail in the south-west of France. Concepts and first results. ACMG Compte rendu d'activité, 4 pp. [Available from J.-F. Berthoumieu, ACMG, Aerodrome de Agen, Agen, France.]
- Bradley, R. A., S. S. Srivastava, and A. Lanzdorf, 1980: Some approaches to statistical analysis of a weather modification experiment. *Statistical Analysis of Weather Modification Experiments*, E. J. Wegman and D. J. DePriest, Eds., Marcel Dekker, 33–53.
- Braham, R. R., 1979: Field experimentation in weather modification. *J. Amer. Stat. Assoc.*, **74**, 57–104.
- Brendler, C. B., and P. C. Walsh, 1992: The role of radical prostatectomy in the treatment of prostate cancer. *CA—Cancer J. Clin.*, **42**, 213–222.
- Bruintjies, R. T., D. W. Breed, G. B. Foote, M. J. Dixon, B. G. Brown, V. Salazar, and J. R. Rodriguez, 1999: Program for the augmentation of rainfall Coahuila (PARC): Overview and design. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 53–59.
- Changnon, S. A., and Coauthors, 1977: Hail suppression: Impacts and issues. *Results from the 1989 Exploratory Cloud Seeding Experiment in Illinois*, Illinois State Water Survey, 427 pp.
- , R. R. Czys, K. R. Gabriel, M. S. Petersen, R. W. Scott, and N. E. Westcott, 1993: *Results from the 1989 Exploratory Cloud Seeding Experiment in Illinois*. Illinois State Water Survey, 154 pp.
- , K. R. Gabriel, N. E. Westcott, and R. R. Czys, 1995: Exploratory analysis of seeding effects on rainfall: Illinois 1989. *J. Appl. Meteor.*, **34**, 1215–1224.
- Cox, C., H. T. Davis, J. F. Calimlim, W. M. Wardell, and L. Lasagna, 1980: Use of the biplot for graphical display and analysis of multivariate pain data in clinical analgesic trials. *Controlled Clin. Trials*, **7**, 63–64.
- Cox, D. R., and D. Oakes, 1984: *Analysis of Survival Data*. Chapman and Hall, 201 pp.
- Davis, B. R., and R. J. Hardy, 1990: Upper bounds for type I and type II error rates in conditional power calculations. *Commun. Stat.*, **A19**, 3571–3584.
- Dennis, A., 1980: *Weather Modification by Cloud Seeding*. Academic Press, 267 pp.
- Epstein, B. E., and G. E. Hanks, 1992: Prostate cancer: Evaluation and radiotherapeutic management. *CA—Cancer J. Clin.*, **42**, 223–240.
- Fleming, T. R., and D. L. DeMets, 1996: Surrogate end points in clinical trials: Are we being misled? *Ann. Intern. Med.*, **125**, 605–613.
- Francis, C. W., V. D. Pellegrini, V. J. Marder, S. Totterman, C. M. Harris, K. R. Gabriel, M. V. Azodo, and M. V. Leibert, 1992: Comparison of Warfarin and external pneumatic compression in prevention of venous thrombosis after total hip replacement. *J. Amer. Med. Assoc.*, **267**, 2911–2915.
- Gabriel, K. R., 1967: The Israeli artificial rainfall stimulation experiment: Statistical evaluation of the period 1961–1965. *Proceedings Fifth Berkeley Symposium on Math. Stat. and Prob.*, L. LeCam and J. Neyman, Eds., Vol. V, *Weather Modification*, University of California Press, 91–114.
- , 1995: Climax again? *J. Appl. Meteor.*, **34**, 1225–1227.
- , 1999: Ratio statistics for randomized experiments in precipitation stimulation. *J. Appl. Meteor.*, **38**, 290–301.
- , and P. Feder, 1969: On the distribution of statistics suitable for evaluating rainfall stimulation experiments. *Technometrics*, **11**, 149–160.
- , and M. Baras, 1970: The Israeli rainmaking experiment 1961–67. Final statistical tables and evaluation. Tech. Rep., Dept. of Statistics, Hebrew University, Jerusalem, Israel, 40 pp. [Available from K. R. Gabriel, Dept. of Mathematics, University of Rochester, Rochester, NY 14627.]
- , and S. A. Changnon Jr., 1982: Piggyback weather experimentation: Superimposing randomized treatment comparisons on commercial cloud seeding operations. *J. Wea. Mod.*, **13**, 7–10.
- , and G. K. Mather, 1986: Exploratory analysis of 1951–1982 summer rainfall data around Nelspruit, Transvaal and of possible effects of 1972–1981 cloud seeding. *J. Climate Appl. Meteor.*, **25**, 1077–1087.
- , and D. Rosenfeld, 1990: The second Israeli rainfall stimulation experiment: Analysis of precipitation on both targets. *J. Appl. Meteor.*, **29**, 1055–1069.
- Gilbert, J. P., B. McPeck, and F. Mosteller, 1977: Statistics and ethics in surgery and anesthesia. *Science*, **198**, 684–689.
- Hills, M., and P. Armitage, 1979: The two-period cross-over clinical trial. *Brit. J. Pharmacol.*, **8**, 7–20.
- Kemphorne, O., 1980: Some statistical aspects of weather modification studies. *Statistical Analysis of Weather Modification Experiments*, E. J. Wegman and D. J. DePriest, Eds., Marcel Dekker, 89–107.
- Levin, Z., Y. Yin, Z. Levin, T. G. Reisen, and S. Tzivion, 1999: Comparison of the effects of hygroscopic and glaciogenic seeding on the evolution of the spectra of cloud and precipitation particles in convective clouds. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 73–74.
- List, R., K. R. Gabriel, B. A. Silverman, Z. Levin, and T. Karakostas, 1999: The rain enhancement experiment in Puglia, Italy: Statistical evaluation. *J. Appl. Meteor.*, **38**, 281–289.
- Mather, G. K., D. E. Terblanche, F. E. Steffens, and L. E. Fletcher, 1997: Results of the South African cloud-seeding experiments using hygroscopic flares. *J. Appl. Meteor.*, **36**, 1433–1447.
- Mielke, P. W., 1985: Design and evaluation of weather modification experiments. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 439–459.
- , 1995: Comments on the Climax I and II experiments including replies to Rangno and Hobbs. *J. Appl. Meteor.*, **34**, 1228–1232.
- Miller, A. J., D. E. Shaw, L. G. Veitch, and E. J. Smith, 1979: Analyzing the results of a cloud-seeding experiment in Tasmania. *Commun. Stat.*, **A8**, 1017–1047.
- Moran, P. A. P., 1959: The power of a cross-over test for the artificial stimulation of rain. *Austral. J. Stat.*, **1**, 47–52.
- Multicenter Diltiazem Post-Infarction Trial Research Group, 1988: The effect of Diltiazem on mortality and reinfarction after myocardial infarction. *New England J. Med.*, **319**, 385–392.
- New York Times, 1999: Sham surgery returns as a research tool. *N.Y. Times Week Rev.*, 25 April 1999, 3.
- Neyman, J., 1977: A statistician's view of weather modification technology (a review). *Proc. Nat. Acad. Sci.*, **74**, 4714–4721.
- Nori, D., 1993: 3-D conformal brachytherapy is lower-cost, outpatient Ca treatment. *Oncol. News Int.*, **2**, 3–22.
- Oakes, D., 1993: The logic and role of meta-analysis in clinical research. *Stat. Methods Med. Res.*, **2**, 147–160.
- , and Coauthors, 1993: Use of compliance measures in an analysis of the effect of Diltiazem on mortality and reinfarction after myocardial infarction. *J. Amer. Stat. Assoc.*, **88**, 44–49.
- O'Brien, P. C., and T. R. Fleming, 1979: A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.

- Orville, H. D., 1996: A review of cloud seeding modeling in weather modification. *Bull. Amer. Meteor. Soc.*, **77**, 1535–1555.
- Piantadosi, S., 1997: *Clinical Trials. A Methodologic Perspective*. John Wiley and Sons, 590 pp.
- Rangno, A. L., and P. V. Hobbs, 1993: Further analyses of the Climax cloud-seeding experiments. *J. Appl. Meteor.*, **32**, 1837–1847.
- , and ———, 1995: Reply. *J. Appl. Meteor.*, **34**, 1233–1238.
- Rosenfeld, D., and W. L. Woodley, 1989: Effects of cloud seeding in west Texas. *J. Appl. Meteor.*, **28**, 1050–1080.
- , and R. Nirel, 1996: Seeding effectiveness—The interaction of desert dust and the southern margins of rain cloud systems in Israel. *J. Appl. Meteor.*, **35**, 1502–1510.
- Schmid, P., 1967: On “Grossversuch III”; A randomized hail suppression experiment in Switzerland. *Proceedings Fifth Berkeley Symposium on Math. Statist. and Prob.*, L. LeCam and J. Neyman, Eds., Vol. V, *Weather Modification*, University of California Press, 141–159.
- Senn, S. J., 1993: *Cross-Over Trials in Clinical Research*. John Wiley and Sons, 296 pp.
- , 1997: *Statistical Issues in Drug Development*. John Wiley and Sons, 423 pp.
- Silverman, B. A., 1999: An independent statistical re-evaluation of the South African hygroscopic flare seeding experiment. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 45–48.
- , W. Sukarmjanaset, D. Rosenfeld, and R. Talumassawadi, 1999: The Thailand warm cloud seeding experiment. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 9–12.
- Steering Committee of the Physicians’ Health Study, 1989: Final report on the aspirin component of the ongoing Physicians’ Health Study. *New England J. Med.*, **321**, 129–135.
- Tukey, J. W., J. R. Brillinger, and L. V. Jones, 1978: Report of the Statistical Task Force to The Weather Modification Board. Vol. II. U.S. Government Printing Office, 94 pp.
- Wegman, E. J., and D. J. DePriest, Eds., 1980: *Statistical Analysis of Weather Modification Experiments*. Marcel Dekker, 145 pp.
- Woodley, W. L., J. Jordan, A. Barnston, J. Simpson, R. Biondini, and J. Flueck, 1982: Rainfall results of the Florida Area Cumulus Experiment, 1970–78. *J. Appl. Meteor.*, **21**, 139–164.
- , D. Rosenfeld, R. Lahav, P. Sudhikoses, N. Tantipubthong, and W. Sukarmjanaset, 1999: The Thailand cold-cloud seeding experiment. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 21–32.
- Yin, Y., Z. Levin, T. G. Reisen, and S. Tzivion, 1999: A numerical evaluation of seeding with hygroscopic flares: Sensitivity to seeding time, seeding height, seeding amounts, size of particles, and environmental shear. Preprints, *Seventh WMO Scientific Conf. on Weather Modification*, Chiang-Mai, Thailand, World Meteorological Organization, 69–74.