Bayesian Methods and Ethics in a Clinical Trial Design
Edited by Joseph B. Kadane

CHAPTER 4

Statistical Issues in the Analysis of Data Gathered in the New Designs

Joseph B. Kadane and Teddy Seidenfeld

The heart of the new designs described in preceding chapters is that the assignment of certain patients to certain treatments is barred by our ethical criteria. Any constraint on the decisions available to a clinical trial comes at a price paid in the statistical efficiency of the trial and possibly the inferences that can be drawn from it. This chapter assesses that price.

There are various philosophies of statistical inference (see Barnett 1982), of which two broad schools will be considered here. The first is a frequentist view, which takes randomization as a necessary element in a design to justify a significance test based on permutations. The second is a likelihood Bayesian view which aims to estimate the magnitudes of important effects. Although often the distinction between likelihood and Bayesian inference is important, the principal argument given here applies simultaneously to both.

4.1 A FREQUENTIST-RANDOMIZATION-SIGNIFICANCE APPROACH

The frequentist-randomization-significance approach has its roots in the writings of Fisher (1971), and is made vivid by the example of the lady tasting tea. She claims to able to tell on the basis of taste, whether the milk or the tea was first put in the cup. To test this claim, Fisher had eight cups prepared, four milk first and four tea first. They were presented to the lady in a random order, but she knew that four were milk-first and four tea-first. If she is able to identify correctly which were milk-first, this would have probability $1/\binom{8}{4} = 1/70$ by Fisher's calculation. Thus, he concludes, an event significant at the level 1/70 has taken place.

Bayesian Methods and Ethics in a Clinical Trial Design, Edited by Joseph B. Kadane
ISBN 0-471-84680-5

This philosophy of experimental design is often used in clinical trials. The idea would be that each patient should be available for assignment to each of the treatments under study. Then the patient is assigned at random, and the outcome is observed. However, adherence to the paradigm poses problems for someone who wishes to adhere to the ethical standard that is the subject of this book. Presumably it would be possible to include in the study only those patients available for randomization to every treatment. This could easily lead to early and inconclusive termination of studies.

If one wishes to take the view that only inference based on randomization are valid, then the ethical premise of this book poses a very sharp dilemma. However, if one takes a less extreme position that other broad streams of statistical thought, such as the Bayesian-likelihood approach, can also be a valid basis for interference, then a reconciliation is possible between the ethical premise and the need to make scientifically supportable analyses. Our current view of the claims of randomization is given in Kadane and Seidenfeld (1990).

4.2 BAYESIAN-LIKELIHOOD APPROACH TO INFERENCE UNDER THE NEW DESIGNS

The argument to be given here is a very general scheme, since it is not specific about exactly which likelihood one might wish to use. That choice depends on the nature of the clinical trial in question. But such specificity is not needed here, for what is said applies to all likelihoods in a broad class.

To begin the argument, we index the patients in the trial by $j = 1, \ldots, J$. Thus we think of patient 1 coming into the trial, being treated, and having an observed outcome, quite possibly before patient 2 comes into the trial. Let X_j $(j = 1, \ldots, J)$ be a vector of observed characteristics of the jth patient, used in deciding what treatment each patient is to receive and possibly other characteristics as well. It is essential to the argument below that whatever characteristics these are, they be measured and known about each patient before that patient is assigned a treatment.

Let T_j $(j = 1, \ldots, J)$ be the treatment assigned to the jth patient, using some design that satisfies the ethical constraint proposed in this book. Similarly, let O_j $(j = 1, \ldots, J)$ be the outcome for the jth patient. This may be a single number for each patient or a vector of outcomes, depending on what is appropriate to the trial in question. For example, one may wish to measure outcome by months of survival, or months of disease-free survival, or by a vector of months of survival of various qualities of life. All those possibilities are permitted by notation O_j.

Also let θ be a vector of the parameters of interest, those that determine the probabilities of outcomes O_j for a patient j given characteristics X_j and treatment T_j. Again this is very general, and it permits a wide range of specifications to address particular clinical trials.

Finally, let $P_j = (O_j, T_j, X_j, O_{j-1}, T_{j-1}, X_{j-1}, \ldots, O_1, T_1, X_1)$ be the past evidence up to and including what is known about patient j. Notice that P_j is

ordered, right to left, in the order that information becomes available: characteristics for patient 1 (X_1), treatment for patient 1 (T_1), outcome for patient 1 (0_1), characteristics for patient 2 (X_2), and so on. When it is necessary to write P_{-1}, this is to be taken as empty.

Using this notation, the likelihood is simply the probability of the data, given θ, which is written $f_\theta(P_J)$. This is the function that is to be analyzed here. The main decomposition used here is solely a consequence of the definition of conditional probability, namely

$$f_\theta(P_J) = \prod_{j=1}^{J} f_\theta(0_j|T_j, X_j, P_{j-1})\, f_\theta(T_j|X_j, P_{j-1})\, f_\theta(X_j|P_{j-1})$$

$$= \prod_{j=1}^{J} f_\theta(0_j|T_j, X_j, P_{j-1}) \prod_{j=1}^{J} f_\theta(T_j|X_j, P_{j-1}) \prod_{j=1}^{J} f_\theta(X_j|P_{j-1}). \quad (4.1)$$

Each of these three products is to be discussed in turn.

The first term $\Pi_{j=1}^{J} f_\theta(0_j|T_j, X_j, P_{j-1})$ is the principle source of information about θ. It is part of the definition of θ as the parameter that

$$f_\theta(0_j|T_j, X_j, P_{j-1}) = f_\theta(0_j|T_j, X_j), \qquad 1 \leqslant j \leqslant J. \quad (4.2)$$

What this means is that θ contains all the information contained in P_{j-1} that might be useful for predicting 0_j from T_j and X_j. Accepting (4.2), the first term can then be rewritten

$$\prod_{j=1}^{J} f_\theta(0_j|T_j, X_j, P_{j-1}) = \prod_{j=1}^{J} f_\theta(0_j|T_j, X_j). \quad (4.3)$$

It should be noted that many clinical trials are examined using a likelihood that amounts to $\Pi_{j=1}^{J} f_\theta(0_j|T_j)$. This suppresses the dependence of the outcome 0_j on patient characteristics X_j and makes it just a function of treatment T_j. We suspect that this practice is unfortunate in general, since there may be important medical mechanisms to be discovered by examining covariates X_j, and the treatment of choice might depend on the covariates X_j. Whether these covariates enter a model as main effects or as interactions, ignoring the covariate information strikes us as risky. However, in the trials of the kind proposed here, an analysis that ignores the covariates X_j, at least those that determine the allocation of treatments to patients, would simply not be legitimate. The additional care in modeling required is one of the important statistical prices to be paid for the new class of designs.

The second term $\Pi_{j=1}^{J} f_\theta(T_j|X_j, P_{j-1})$ has to do with how treatments are assigned. With all the designs discussed here, T_j, the treatment assigned to the jth patient, is a known function of X_j, that patient's characteristics, and the past history P_{j-1}. In many trials it is deterministic, meaning that for each configuration (X_j, P_{j-1}) one of the treatments has probability one of allocation and all the others probability zero. If randomization were part of the design, which is definitely permitted provided that the ethical constraint is maintained,

two treatments might have probability (1/2) each, or three have (1/3) each, and so on. Other, unequal, probabilities might also be used. But the key point is that in none of these design does $\Pi_{j=1}^{J} f_\theta(T_j|X_j, P_{j-1})$ depend on θ. Hence as far as the likelihood function is concerned, this second factor does not enter.

One of the most hotly debated matters among the researchers for this book is the question of whether patients may choose a treatment for themselves. The proponents of this view urged that patients be allowed access to the full data set, and whatever advice they wished, and then be allowed to choose their treatment. This idea was resisted by those who felt that the result would be a serious degradation of the scientific quality of the data. The impact of that determination comes in the examination of this second term. If patients were to choose their own treatments, the argument that $\Pi_{j=1}^{J} f_\theta(T_j|X_j, P_{j-1})$ does not depend on θ would fail. It would now be necessary to explain statistically the behavior of patients in choosing their treatments, and there might well be contamination between these choices and the effect of the treatment itself.

As a simple example, there was recently a clinical trial of treatments for breast cancer in which one treatment involved lumpectomy, removal of only the tissue immediately surrounding the tumor, in comparison with the more traditional radical mastectomy, which removes the entire breast (Fisher et al. 1985). Suppose that patients were permitted to choose between these two operations. It is entirely possible that women would differ in their choices for reasons having complicated interactions with the outcome of the treatment itself. For instance, such choices might be correlated with social class or with suspected severity of disease. It is certainly possible to report the findings of such a trial in terms such as "Of the women who chose lumpectomy, the five-year survival rate was x, while of those who chose radical mastectomy, the survival rate was y." But data reported this way are not directly useful to a new patient or her attending physician. The best advice would be "Be like those that chose lumpectomy," if $x > y$. This is difficult advice to act upon because the trial cannot disentangle the effect of the treatment itself from the effect of the sort of patients who chose it.

Medical science has been confounded so many times on this very point that we feel that caution is warranted. One famous historical example, that of the Lanarkshire milk experiment, is reviewed in the next section.

We do wish to record, however, that we find the idea of patient choice appealing from an ethical and legal perspective. With the technology available to us now, however, we do not see how to allow patients to choose their treatment and not confound those choices with treatment effects. Simply that we do not now see how is not to prove that such a thing cannot be done.

Returning to the factorization of the likelihood in equation (4.1), we now discuss the third term $\Pi_{j=1}^{J} f_\theta(X_j|P_{j-1})$. This term will not depend on θ provided that the kinds of patients that are proposed for, and agree to be in, clinical trials are not a function of past patient outcomes. We believe that the kind of trial we propose avoids such functional dependence because we make use of past history P_{j-1} for the benefit of patients, obviating their need to do

so for themselves, which would affect the likelihood through this third term. We do not know the extent to which data contamination of this kind affects standard clinical trials.

In summary, the discussion here leads to the conclusion, in a single equation, that for the trials here considered,

$$f_\theta(P_J) \propto \Pi_{j=1}^{J} f_\theta(0_j | T_j, X_j). \tag{4.4}$$

This is the form that we use to evaluate the results of a clinical trial of the kind considered in this book.

4.3 THE LANARKSHIRE MILK EXPERIMENT AND "BIASED" ALLOCATIONS

An illustration helps to point out the importance of the requirement in experimental design that probabilities for treatment allocations are to be a known function of recorded patient characteristics. When this design feature is not present, that is, when allocations are based on unrecorded patient characteristics or other factors, the experiment is made susceptible to undetectable and uncorrectable biases. The Lanarkshire milk experiment offers a vivid example of this difficulty.[1]

For four months early in 1930, the Lanarkshire school district of Scotland conducted a large scale nutritional experiment to test the value of a 12-oz. daily milk supplement. The experiment involved 20,000 students, ranging in ages between 5 and 11. Half of these, the "feeders," received 3/4's of a pint of grade A (Tuberculin tested) milk during school days, and the remaining 10,000 used as "controls," received no additional milk in their diet. Moreover the "feeders" were divided equally between those who received raw milk and those who received pasteurized milk. Thus, in all, there were three treatment groups: raw milk "feeders" ($T1$), pasteurized milk "feeders" ($T2$), and "controls" ($T3$), in the ratios of 1:1:2.

The trial commenced in February 1930 and ended in June of that year. According to the planned design, "feeders" and "controls" were to have been randomly chosen from 67 schools. Between 200 and 400 students were selected from each school, with half receiving milk and half not. For simplicity of the administration, each of the 67 schools had only one variety of feeder: 34 schools allocated raw milk and 33 allocated pasteurized.[2]

For each of the 20,000 students a record was made of age (by year), sex, and both pre-trial and post-trial weights and heights. That is, the design planned 14 categories of students: 7 (for age) $\times$ 2 (for sex), with a roughly 1:1:2 randomized allocation of three treatments within each category. The nutritional benefits of the rival treatments were analyzed by a contrast of growth rates, using differences in the initial and final heights and weights, with students blocked according to sex and age.

Leighton and McKinley's (1930) official report proposed three findings about the milk supplements, which we quote:

[1] The influence of milk to the diet of school children is reflected in a definite increase in the growth both in height and weight.

[2] There is no obvious or constant difference in this respect between boys and girls and there is little evidence of definite relation between the age of the children and the amount of the improvement. The results do not support the belief that the younger derived more benefit than the older children. As manifested merely by growth in weight and height, the increase found in younger children through the addition of milk to the diet is certainly not greater than, and is probably not even as great as, that found in older children.

[3] Insofar as the conditions of this investigation are concerned, the effects of raw and pasteurized milk on growth in weight and height are, so far as we can judge, equal.

Addressing these claims in reverse order, it should be noted that Fisher and Bartlett (1931) argue for a definite advantage in raw over pasteurized milk. Regarding the thesis that there is little connection between age and the nutritional benefits of the additional daily milk, a condensed version of some summary statistics, reproduced in Table 4.1 taken from *Student*'s (1931, p. 403) discussion, suggests otherwise. As Gosset notes, the observed weight gains, at least, are impressively larger in older children than in younger ones, whereas there is some evidence of the reverse trend for boys' heights and no clear indication of the relevance of age in predicting the contribution of the added milk to girls' heights.

The principal difficulty with the Lanarkshire study, however, becomes apparent in a more detailed account of the treatment allocation rule employed. Since the official statistical evaluation of the rival treatments rested on the observed differences between pre-trial and post-trial measurements of heights

Table 4.1. Summary statistics for Lanarkshire milk experiment

	Weight gains by "feeders" over "controls"[a]		Height gains by "feeders" over "controls"[b]	
Ages (yrs)	Boys	Girls	Boys	Girls
5, 6, and 7	1.13 ± 0.73 oz 9%	1.23 ± 0.72 oz 13%	0.083 ± 0.011 in. 11%	0.059 ± 0.011 in. 8%
8 and 9	3.15 ± 0.68 30	4.47 ± 0.67 51	0.071 ± 0.011 10	0.098 ± 0.010 14
10 and 11	5.21 ± 0.85 78	7.88 ± 0.79 73	0.037 ± 0.012 5	0.055 ± 0.012 8

[a] In oz. and in % "control."
[b] In in. and in % "control."

and weights, any departure from the planned random allocation rule to a rule which, instead, tied a particular treatment to an independent propensity for measured growth, inadvertently biased the study by convoluting that propensity with the nutritional benefits of the milk. More generally, this problem occurs whenever there is correlation between an outcome, however measured, and the division of subjects under the allocation rule.

Using simple randomization[3] we have the known allocation probabilities for each student j ($j = 1, \ldots, 20{,}000$) in the experiment:

$$\text{Prob}\,(T_{i,j} | \text{age}_j, \text{sex}_j, \text{initial height}_j, \text{initial weight}_j) = 0.25, \qquad i = 1, 2$$
$$\text{Prob}\,(T_{3,j} | \text{age}_j, \text{sex}_j, \text{initial height}_j, \text{initial weight}_j) = 0.5.$$

However, to understand how treatments were assigned, Gosset quotes from the Leighton and McKinley report as follows:

> The teachers selected the two classes of pupils, those getting milk and those acting as 'controls,' in two different ways. In certain cases they selected them by ballot and in others on an alphabetical system.... In any particular school where there was any group to which these methods had given an undue proportion of well fed or ill nourished children, others were substituted in order to obtain a more level selection.

Thus the way was opened for teachers to make treatment allocations based on all sorts of considerations unrecorded in the study. The upshot of this unregulated freedom was, not surprisingly, a "control" group with noticeably larger initial heights and weights. In the opinion of the official report, the "controls" were found to have about a three-month growth advantage in weight and about a four-month advantage in height over the "feeders." Given the opportunity to reallocate treatments after the randomized allocations were determined, the speculation is that, perhaps subconsciously, the teachers followed their sentiments to do what was in the best interests of their students and provide free milk to those who stood most in need of it!

Let us suppose, then, that the allocation rule employed was, qualitatively, of the sort agreeing with the observed differences between "feeders" and "controls." Suppose that it was an allocation geared to promote the welfare of the subjects:

$$\text{Prob}\,(T_{i,j} | \text{age}_j, \text{sex}_j, \text{low initial height}_j, \text{low initial weight}_j) > 0.25, \quad i = 1, 2,$$

$$\text{Prob}\,(T_{3,j} | \text{age}_j, \text{sex}_j, \text{low initial height}_j, \text{low initial weight}_j < 0.5.$$

How might this alter the experimental findings?

If, as hypothesized, the "controls" were on average better nourished than the "feeders" (despite 3/4 pint of milk added daily to the latters' diet), then treatments were convoluted with other factors promoting growth in such a way as to *mask* the nutritional benefits of the milk. How then are we to explain the

remarkable weight gains to the "feeders," as summarized in Table 4.1. The answer lies with the technique for measuring gains!

Students were weighed and measured in their indoor clothing. In February, for the pre-tests, they were dressed in cold weather attire and in June, for the post-tests, they wore springtime garb. If, as suspected, the "controls" were from homes that could afford better diets, were they not also from homes that could offer heavier winter clothes? Add to this Gosset's observation that smaller children have more limited wardrobes, permitting fewer discards. In effect, the "controls" shed more weight between the pre- and post-trial measurements than did the "feeders," and that in proportion to their age. A graph of the growth curves for the six groups: 3 (for treatments) × 2 (for sex), confirms this speculation and solves the mystery of the miraculous effects of an extra 0.75 pint/day of milk on the weight gains of the older children; see Figures 4.1 and 4.2.

In conclusion, we see that even a large-scale study, on the order of the Lanarkshire milk experiment, is susceptible to "bias" when treatment allocations are made according to unrecorded factors. In the Lanarkshire case, by permitting teachers to choose treatments in order to maximize the welfare of their students and making no statistically useful note of that fact, the evaluation of the rival treatments (on the false supposition of the allocation rule originally intended) served, ironically, to overestimate the beneficial effects of the milk supplements. What was faulty with the experiment was not so much the lack of a randomized allocation of treatments, though in all probability that would have avoided the pitfall encountered. Rather, the official report was left without the basis for statistically cogent analysis of the experimental data when, too late, it was discovered that allocations were made according to individual assessments of unrecorded factors. Once that was permitted, there was no basis in the data for a rebuttal to the familar skeptical challenge, most appropriate in this case, to wit: the charge that the study is "biased" for convoluting treatments with other (independent) causes of the effects under investigation.

Had the allocation been, instead, by a rule under which the probabilities of assignment to treatment were known, the experimenters could have responded as follows:

> Your skepticism requires the added doubt that other important (independent) causes are convoluted with treatments in ways that cannot be accounted for by the recorded values of the variates that we used to determine the treatments. Since we have taken care to make the allocations a known function of the recorded factors, the very factors that we think are important to the effects under investigation, your skepticism is without basis in fact.

Of course that is exactly what the investigators in the Lanarkshire study could not say.

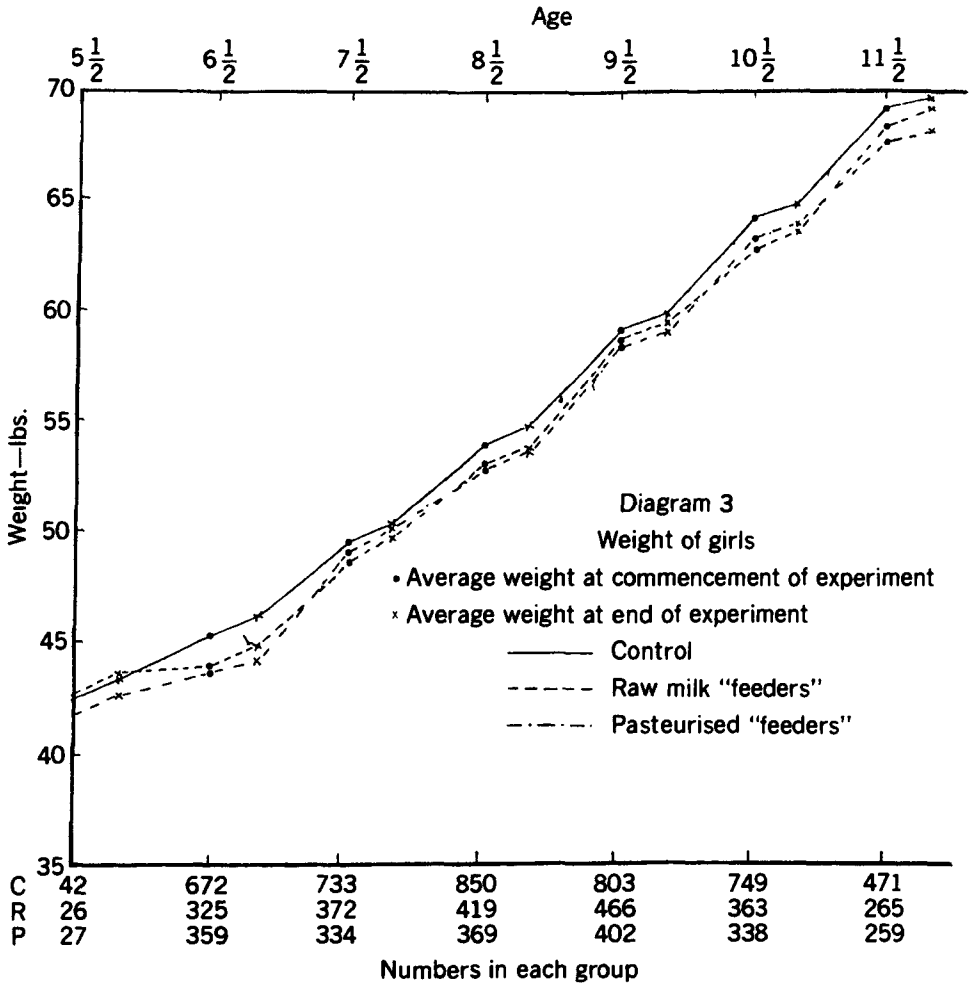

Figure 4.1

4.4 CONCLUSION

This chapter has served two purposes. First, we developed the assumptions leading to the conclusion expressed in (4.4), which in turn leads to reasonably simple analyses of clinical trials of the class considered here. The principle statistical price paid for the ethical constraints turns out to be that we must explicitly condition on the patient characteristic X_j.

Second, we exposed at some length the reasons for caution in controlling what patients get which treatments in a clinical trial. This is discussed briefly in Section 4.2, and at greater length in terms of an example in Section 4.3. This

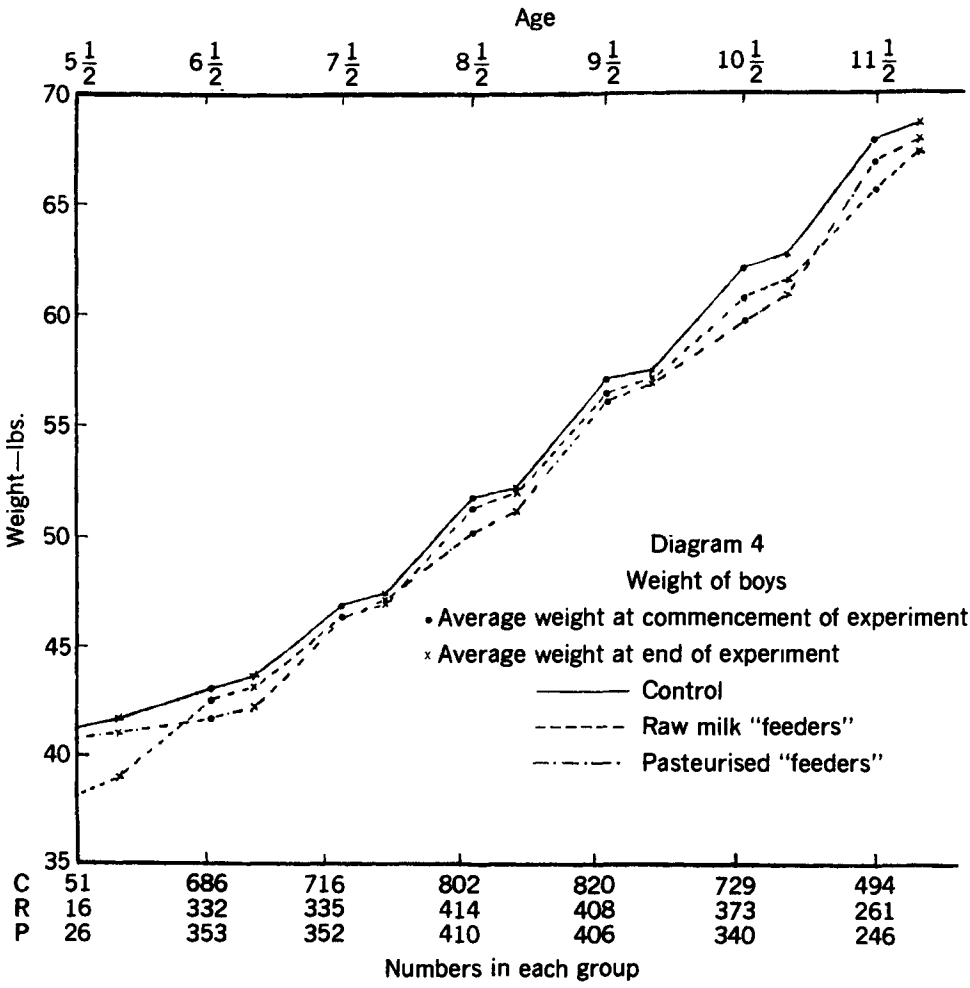

Figure 4.2

caution leads us in the direction of recommending a controlled trial of the kind we do, rather than an uncontrolled, unrestricted patient-choice trial.

NOTES

[1]We rely on Gosset's *Students* 1931 summary of this experiment.

[2]The simple precaution for securing a reliable separation of the two "feeder" groups, limiting each school to one variety of milk, comes at the expense of making questionable an hypothesis used for the statistical analysis within the official report. The issue thus raised (Gosset, p. 399) was whether the "controls" were suitably homogeneous to

warrant comparisons of the combined "control" averages from all 67 schools with the two "feeder" averages taken from only half of the schools.

An unbalanced division of the 67 schools between the two kinds of feeders, according to health or socioeconomic status, opens the door to "Simpson's" paradox. See Lindley (1983) for a good discussion of this problem. However, inspection of the growth curves for the "feeders" (diagrams 1–4, pp. 400 and 402) does not lend credence to this worry in this case.

[3]A simple randomization ignores the potential bias associated with an imbalance between the two groups of "feeder" schools, discussed in note 2.

REFERENCES

Barnett, V. (1982), *Comparative Statistical Inference*, New York: Wiley.

Fisher, R. A. (1971), *The Design of Experiments*, 9th ed., New York: Hafner Press.

Fisher, B., Bauer, M., Margolese, R., Poisson, R., Pilch, V., Redmond, C., Fisher, E., Wolmark, N., Dentsch, M., and Montague, E. (1985), "Five year results of a randomized clinical trial comparing total mastectomy and segmented mastectomy with or without radiation in the treatment of breast cancer," *New England Journal of Medicine*, **312**, 665–673.

Fisher, R. A., and Bartlett, M. S. (1931), "Pasteurized and raw milk," *Nature*, **127**, 591–592.

Gosset, W. S. (1931), "The Lanarkshire milk experiment," *Biometrika*, **23**, 398–406.

Kadane, J. B., and Seidenfeld, T. (1990), "Randomization in a Bayesian perspective," *Journal of Statistical Planning and Inference*, **25**, 329–345.

Leighton, G., and McKinley, P. L. (1930), *Milk Consumption and the Growth of School Children*, London: H.M. Stationery Office.

Lindley, D. (1983), *The Role of Randomization in Inference*, in P.S.A. 1982: Procedings of the 1982 Biennial Meeting of the Philosophy of Science Association, P. D. Asqueth and T. Nickels (eds.), vol. 2, East Lansing: Philosophy of Science Association, pp. 431-446.