Drug Information Journal, Vol. 35, pp. 1479–1494, 2001
Printed in the USA. All rights reserved.

0092-8615/2001

INDIVIDUAL THERAPY: NEW DAWN OR FALSE DAWN?

STEPHEN SENN, BA, MSC, PHD

Professor of Pharmaceutical and Health Statistics, Departments of Epidemiology and Public Health and Statistical Science, University College London, London, United Kingdom

The sequencing of the human genome brings with it the hope that greater understanding of genetic components of disease will allow the more specific targeting of therapies. It has also been suggested that it will permit sponsors to run "cleaner" clinical trials with less variability and a consequent saving in patient numbers. However, we do not know how much of the variation in response that we see from patient to patient in clinical trials is genetic, because we rarely design the sort of trials that would allow us to identify patient-by-treatment interaction. Such interaction provides an upper bound for gene-by-treatment interaction for a group of patients studied since patients differ by more than their genes. On the other hand, however, the variability seen within a clinical trial may generally be expected to be less than the total variation that would be seen within a population. There is a related statistical issue to do with the interpretation of effects from clinical trials. This arises because there is confusion between experimental and sampling models of clinical research. It is concluded that we may have to pay careful attention to certain design features of clinical trials if we wish to make progress in this field.

Key Words: Patient-by-treatment interaction; Cross-over trials; n-of-1 trials; Effect sizes

INTRODUCTION

"IT WILL SOON be possible for patients in clinical trials to undergo genetic tests to identify those individuals who will respond favourably to the drug candidate, based on their genotype, and therefore the underlying mechanism of their disease. This will translate into smaller, more effective clinical trials with corresponding cost savings and ultimately better treatment in general practice. In addition, clinical trials will be capable of screening for genes involved in the absorption, metabolism and clearance of drugs and the genes which are likely to predispose a patient to drug-induced side-effects. In this way, individual patients will be targeted with specific treatment and personalised dosing regimens to maximise efficacy and minimise pharmacokinetic problems and other side-effects." (1)

The writer is Sir Richard Sykes, FRS, at the time chief executive officer of Glaxo-Wellcome and now chairman of GlaxoSmith-Kline and rector of Imperial College London. The statement thus deserves attention. Nevertheless, it will be claimed here that the hope expressed in the cited passage, and which has been expressed elsewhere by others (2), may be rather more difficult to realize than has been supposed.

This argument of this paper is organized as follows. First, some general statistical

Presented at the DIA Workshop "Statistical Methodology in Clinical R &D," April 2–4, 2001, Vienna, Austria.

Reprint address: Stephen Senn, BA, MSc, PhD, Professor of Pharmaceutical and Health Statistics, Department of Epidemiology and Public Health and Statistical Science, University College of London, London, United Kingdom. E-mail: stephens@public-health.ucl.ac.uk.

points are made about variation in clinical trials. Next, three published examples, the first an invalid analysis of a multicenter trial, the second an invalid proposal for analyzing clinical trials, and the third an inappropriate analysis of individual response in a trial that could have validly identified it, are considered to illustrate the claim that variability in clinical trials is often misunderstood. In these three examples patient-by-treatment interaction is assumed to apply without a valid demonstration that it exists. A further, positive example of exactly the sort of trial that would be capable of identifying patient-by-treatment interaction is then discussed. Some lessons are then drawn as to what steps might be taken to identify patient-by-treatment and then gene-by-treatment interaction. Finally, a related statistical issue to do with variation in clinical trials, and in particular whether sampling or experimental inference is appropriate, will be mentioned.

VARIATION IN CLINICAL TRIALS

The argument that will be made here generalizes easily to trials with many treatments but is conveniently discussed in terms of a two-armed trial. Accordingly, this is assumed to be the case in what follows. The variation seen in randomized clinical trials is conveniently divided into the following four major sources, not all of which will necessarily be identifiable, depending on design. First we have the "main effect" of treatment, that is to say, the average difference over all randomizations between the outcome (measured on some suitable scale) under the experimental treatment and the outcome under control. Second, we have the "main effect" of patients, that is to say, the general difference that might be assumed to exist between a group of patients under homogenous treatment, whether that treatment is the control or the experimental preparation. Third, we have patient-by-treatment interaction. This can be regarded as being the variability introduced into clinical trials by virtue of some patients responding more favorably to a given treatment than others. It implies that the difference that is made by giving a patient the experimental treatment rather than the control treatment varies from patient to patient. Finally, we have within-patient error. This is the extent to which the effect of the same treatment given to the same patient might vary from occasion to occasion. This effect includes not only random variation in the state of the patient but also any uncontrollable measurement error in the trial. These effects are summarized in Table 1.

The descriptions of these various effects already contain hints as to the circumstances under which they are separately identifiable. For example, consider a parallel group trial. Since we do not measure the same patient repeatedly under the same treatment we can hardly identify D, the within-patient error. Nor, since we do not treat each patient with each treatment, as say, in a cross-over trial, can we identify C, the patient-by-treatment interaction. This is not to say that these factors are not present in such a trial; it is just that the effects C and D cannot be separated from B. The patients within one treatment arm will differ in their response not only because they would differ from each other whatever treatment they were all given (B) but also because some are currently experiencing temporary difficulties which others are not (D) and also because some but not all are responding poorly to *this* treatment (C). What we observe is the joint effect of B, C, and D. They cannot be separated. The consequences of nonidentifiability for various types of clinical trial are given in Table 2.

The table, while summarizing the general position, is somewhat of an oversimplification. For example, if in a parallel group trial we can divide patients into males and females, we can then separately resolve a portion of B, the main effect of patients into the main effect of sex: part of the difference observed between patients is the difference between the sexes. We can resolve a portion of C into sex-by-treatment interaction, since we can estimate separately the effect of treatment for both males and females. The position then is, that although a *total* separation

TABLE 1
Sources of Variation in Clinical Trials

Label	Source	Description
A	Between treatments	The average difference between treatments over all randomizations (and hence over all patients). The 'true' mean difference between treatments
B	Between patients	The average difference between patients. (Averaged over both experimental and control treatments.)
C	Patient-by-treatment interaction	The extent to which the difference between treatment differs from one patient to another. (Equivalently, the extent to which the difference between patients being given the same treatment depends on treatment given.)
D	Within-patient error	The variability shown from treatment period to treatment period when the same patient is given the same treatment

of all of C from all of B is not possible, a separation of part of C from part of B is.

Similarly, if on completing a parallel group clinical trial, we find a much higher total variability in one arm than the other, a possible explanation is that there is a subgroup of patients who are responding to treatment in one arm (3). Again a partial (but very imperfect) identification of C becomes possible.

Finally, if we have repeated measures within a given treatment period, then it may be possible to fit a random effects model to the data (for example, a random slopes model) and thus identify treatment-by-patient interaction. This approach is commonly used by the population pharmacokinetics school and can be extremely powerful (4). Because the repeated measures occur with treatment periods rather than, as in the case of repeated measures cross-over, over periods to which treatment has been randomly assigned, such approaches do, however, require fairly strong modeling assumptions to succeed. For example, lack of fit for a given model will contribute (inappropriately) to the error term D.

In general, however, the situation is roughly as indicated in Table 2. Patients must be measured on more than one treatment to identify between-patient error. They must be measured more than once on each treatment to identify treatment-by-patient interaction. If the logic of the randomized clinical trial is to be fully exploited, this extra measurement must come through extra periods with random assignment rather than measures within periods.

Now consider the implication this has for the prospect that genotyping will make clini-

TABLE 2
Identifiability and Clinical Trials

Type of Trial	Description	Identifiable Effects	Error Term
Parallel	Each patient receives one treatment	A	B + C+ D
Cross-over	Each patient receives each treatment in one period only	A and B	C + D
Repeated period cross-overs (sets of n-of-1 trials)	Each patient receives each treatment in at least two periods	A and B and C	E

cal trials much cleaner and reduce the variability we see in them. A factor that affects the magnitude of B will be genetic variation between patients. A very simple example can be given in trials of asthma. Other things being equal, taller individuals tend to have higher forced expiratory volume in one second (FEV_1) than shorter ones. Since height is at least partially genetically determined, this is very plausibly an example of (partial) genetic determination of FEV_1. However, it is also obviously of no practical interest. Since we can measure a patient's height we do not need to measure his or her genetic inheritance as regards height. We can use height itself as a covariate in analyzing the clinical trial. Of course, one might argue that this is a trivial example. There are many genes whose effects appear hidden; by identifying these we can achieve a useful stratification of patients that will allow us to eliminate a part of the variation currently assigned to the error term.

However, to the extent that the main effect of genes is relevant to the outcome in many trials it is also likely to be relevant to the baseline measurements in these trials. Suppose that there is a genetic subgroup of extreme asthmatics. We shall see some asthmatics with lower baseline values than others. In using these baselines in an analysis of covariance (as is now common within the pharmaceutical industry, although elsewhere practice lags behind) we have already largely taken account of this genetic effect without having had to identify it. Of course, baseline values are measured with some error so that it is not true that the partial regression of outcome on gene given baseline will be zero. Nevertheless, given our ability to measure and use predictor covariates (including repeated baselines) in clinical trials we should not assume that the further ways that genotyping will deliver will make much of a "main effect" contribution to reducing variability in clinical trials.

If a contribution can be made to reducing variability, it will be in our ability to reduce the size of the C term in clinical trials by restricting entry to individuals whom we could identify with the help of these new techniques as being "responders." In other words, it is the contribution of gene-by-treatment interaction to the patient-by-treatment interaction term that is important. This seems to be what Sir Richard Sykes was discussing in the passage quoted.

However, for such reduction in variability to make a useful contribution to reducing variability in clinical trials, two conditions are necessary. First gene-by-treatment interaction has to be important. In this connection, it is not only necessary for there to be relevant genetic differences between patients for this to make an appreciable contribution to overall variance, the frequency of relevant varying minority subgroups also has to be large enough (5, Chapter 4). Second, we have to be capable of finding such variation. The second point will not be considered further in this paper except to note that unless it is the case that such variability in response is attributable to a few alleles acting independently, the identification of responders and nonresponders may prove extremely difficult (5). The rest of this paper is concerned instead with the first point. What evidence is there that gene-by-treatment interaction has been (hitherto) a hidden and important source of variability in clinical trials?

It is here that the concept of patient-by-treatment interaction becomes important. Patient-by-treatment interaction provides an upper bound on gene-by-treatment interaction. This is because patients differ not only by their genes but by their environmental circumstances and histories including a whole host of factors that are not completely genetically determined (although some of them may have a genetic component). Such factors that diet, exposure to pathogens, social circumstances (for example, whether married or single), habits (for example, smoking, drinking, sexual activity, degree of exercise taken), income, occupation and so forth. Thus, it would seem plausible that unless patient-by-treatment interaction is large, gene-by-treatment interaction cannot be so. The point is similar to one that has been made recently by Elston in an extremely useful introduction to statistical methods in genetic epidemiology (6). He writes, of an analogous phenomenon, "The presence of familial aggregation is no guarantee that there is a genetic component to the aetiology of a trait, but if there is no familial aggregation there is no point in search for segregating loci that might cause such an aggregation." (6, p. 532)

If, however, we study Table 2, we see that the only sort of trial capable of identifying patient-by-treatment interaction is the repeated period cross-over in which patients are randomized to sequences of treatments such that each treatment is received in more than one period. Such designs, however, are very rarely employed. Ironically, one of the few fields in which they have been employed is that of bioequivalence, where (since two formulations of the same drug are being compared) important subject-by-treatment interaction is inherently rather implausible. The question then arises, what evidence do we have that the reason that we see substantial variability in clinical trials is due to patient-by-treatment interaction? Before considering this point we look at three examples.

PATIENT-BY-TREATMENT INTERACTION? THREE EXAMPLES

The first example concerns a reanalysis of the beta-blocker in heart attacks trial (BHAT) by Horwitz et al. (7). This trial compared propranolol to placebo using 3837 male and female patients aged 30 to 69 who had been hospitalized with an acute myocardial infraction in one of 31 clinical centers. The minimum follow-up was 12 months and the mean was 25 months. There was an overall benefit in terms of mortality for propranolol (7.2%) compared to placebo (9.8%) with an estimated unadjusted odds ratio (95% confidence limits) of 0.71 (0.57, 0.90) in favor of propranolol. (Because Horwitz et al. do not give the numbers on each treatment group it is not possible to perform this calculation precisely.)

Horwitz et al. then noted that although the mortality rate was lower under propranolol than placebo in 21 centers (which they named dominant), there were 10 centers (which they named divergent) in which the survival rate was better under placebo. The situation is illustrated in Figure 1, which plots center by center the mortality rate under propranolol against that under placebo. The line is the line of exact equality. They also compared divergent and dominant centers for baseline characteristics using the chi-squared test on one degree of freedom. A summary of some of their findings is given in Table 3. Some factors were apparently highly significantly different between centers, for example, although the critical value at the 5% level for a chi-square with one degree of freedom is 3.84 the value for race white/other was 95.5.

They then compared the divergent and dominant centers regarding mortality using the Gail-Simon test (8), finding a significant difference between the two groups of centers. Their conclusions were: "This new approach to the analysis of multicenter trials generally, and to the recognition that treatment can have widely different effects for some patients specifically, promises to make trial results helpful both to regulatory agencies who license drugs, and to physicians and their patients who must use the drugs."(7, p. 400)

However, the "new approach" is invalid on several accounts and will not be accepted by regulatory agencies, which are much more vigilant, when it comes to statistical analysis, than either journal editors or referees (9,10). The paper by Horwitz et al. is, unfortunately, a demonstration of the growing gulf between standards inside and outside of the pharmaceutical industry, to the detriment of the latter.

There are a number of serious errors in this approach. First, a simple chi-square cannot be used to compare centers as regards demographic factors because of probable clustering of factors by center. A chi-square analysis of a 2 x 2 table assumes independence of all observations within any one of the four cells. However, people tend to cluster so that this independence does not apply and it is hardly surprising, for example, that

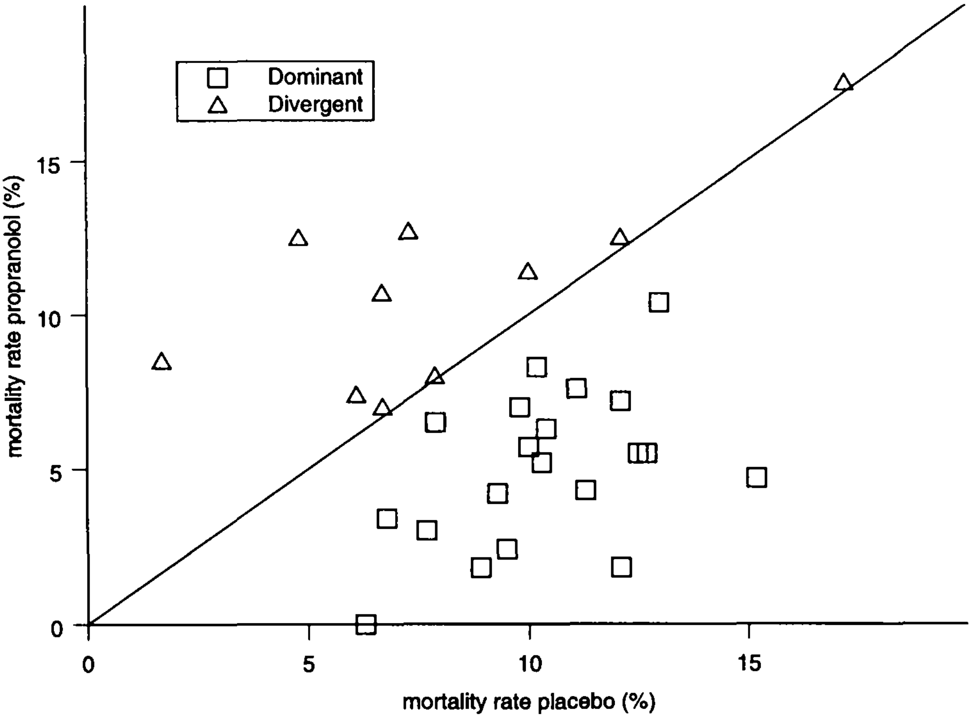

FIGURE 1. Mortality rates for propranolol against placebo for 31 centers in the BHAT trial.

race was the factor that produced the highest chi-square. Second, the authors used the Gail-Simon test in a way that its authors specifically forbade in that they made a post-hoc selection of centers on the basis of results (8). Despite themselves warning that their approach would only be valid if, "a formal test for qualitative interaction is performed that shows a significant effect that excludes chance as an explanation for the divergent results" (7, p. 399), Horwitz et al. were unrepentant (11) when it was pointed out to them that their approach used a test in a way that was formally forbidden (9). Finally, when a valid analysis of these centers is performed, this reveals an astonishing fact. The degree of variation between centers of the treatment effect is slightly less than one might expect by chance (9,10).

This is illustrated by Figure 2, which gives a Galbraith plot of the results (12,13, 14). The Y-axis gives the standardized estimate, the ratio of the log-odds ratio to its standard error for each of the 31 centers. The X-axis gives the precision, the reciprocal of the standard error. The slope of regression through the origin of these points gives the overall treatment estimate for the trial. The fact that it slopes slightly downwards indicates the modest benefit for propranolol found in this trial. The lines that parallel this slope are at ±1.96 standard estimates. One

TABLE 3
Distribution of Some Factors in the BHAT Trial

	Dominant		Divergent		
	No.	%	No.	%	χ^2
White	2113	85.2	1297	95.6	95.5
Prior MI	1099	44.3	509	37.5	16.7
History of CHF	249	10.2	102	7.6	6.7

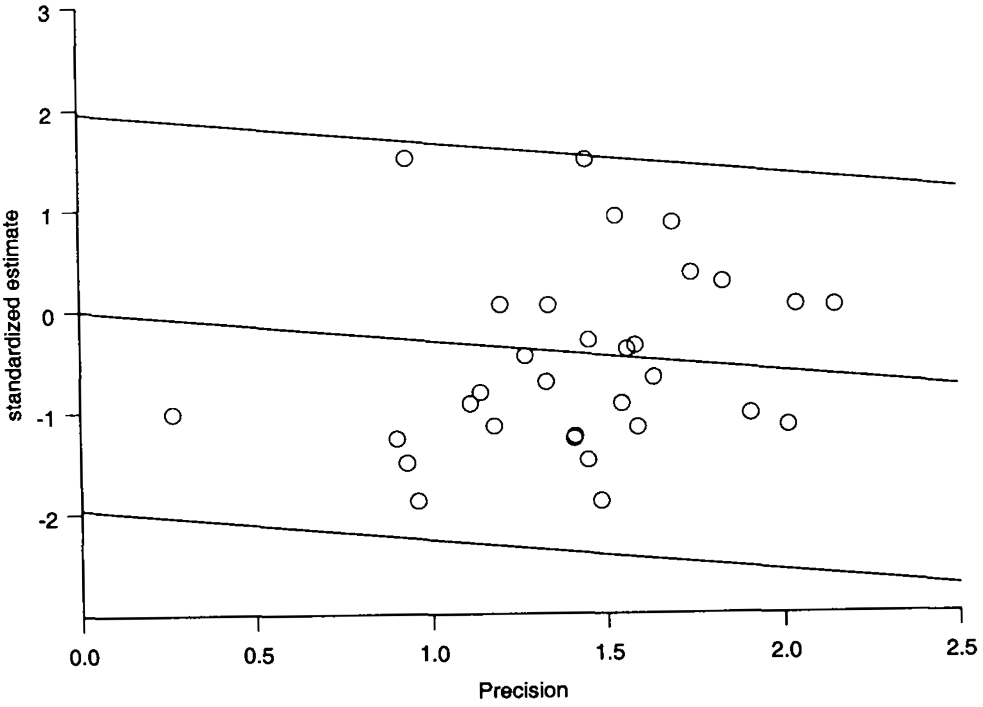

FIGURE 2. Galbraith plot of the 31 centers in the BHAT trial. The slope of the lines is the fixed effects estimate of treatment effect. The outer lines are at plus and minus 1.96 standard errors.

in 20 points would be expected to lie outside these limits by chance; none of them do and this is consistent with a random effects analysis for this trial using the method of Hardy and Thompson (15), which produces an estimate of the random effect that is a little below zero (9, 10). Thus, the data do not give the slightest reason for supposing that the true effect of treatment was not identical in every center.

It is, unfortunately, all too easy to underestimate the effect of chance in multicenter trials. If we have designed a trial with 80% power for a 5% level of significance (two-sided) and the clinically relevant difference obtains, we only need six centers before it is odds-on that at least one of them will show an "effect reversal," that is to say, an apparent superiority for placebo (16). Figure 3 shows a number of probability distributions for the number of effect reversals for trials with 80% power. Each distribution corresponds to a different number of centers. It can be seen that for 30 centers, 9 is the most likely number of effect reversals. Of course, the power of many trials may be greater than 80%. However, even with 99% power, an unusually high value for power, the probability of no effect reversals in a 31-center trial is less than one in 2000. This can be calculated simply as follows. The percentage point for 99% is 2.3263 and that for 2.5% is 1.9600. The sum of these is 4.2863 and is thus the noncentrality parameter for the trial. However, if the 31 centers are of equal size the standard error in each center is $\sqrt{31} = 5.5678$ times as large as for the trial as a whole. The noncetrality parameter is thus $4.2863/5.5678 = 0.77$. The probability that there is no effect reversal in a given center is then found easily from tables of the Normal distribution to be 0.78. The probability of no effect reversal in any center is then $0.78^{31} \approx 1/2200$ In other words, had no effect reversals been found in

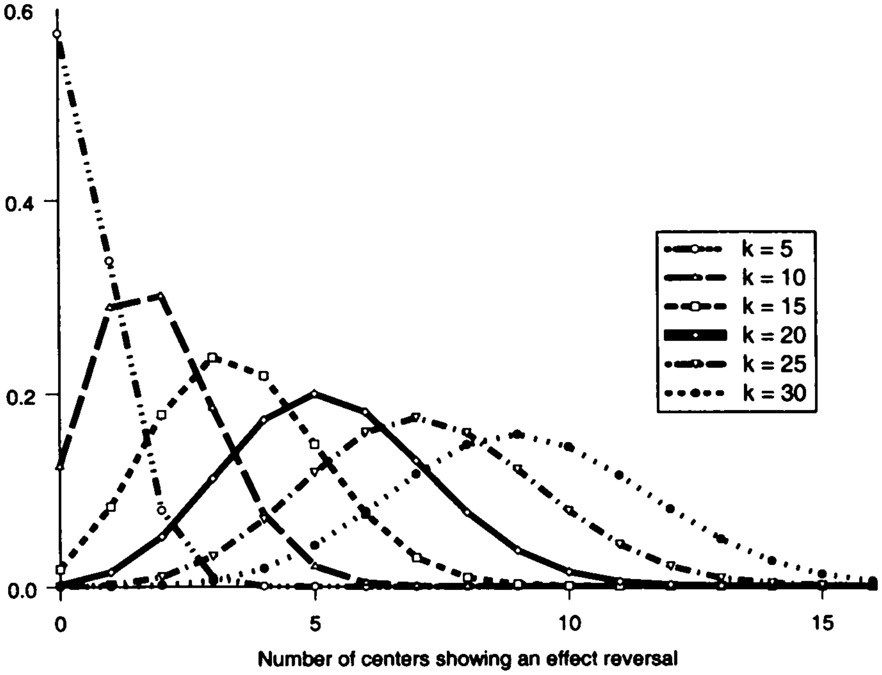

FIGURE 3. Probability distribution of the number of centers showing an effect reversal for trials with varying numbers of centers, k. It is supposed that the trials have been designed with 80% power for a 5% type I error rate and that the clinically relevant difference obtains.

the BHAT trial, there would be good grounds for suspecting fraud.

The second example is a paper by Guyatt et al. proposing an analysis of clinical trials that they suggest will permit one to identify the proportion of patients showing a benefit (17). They consider the specific example of a cross-over trial comparing salmeterol to salbutamol to placebo in asthma as regards quality of life. Because each patient has been treated with each treatment, it is possible to calculate a pair-wise difference for each treatment for each of the three possible pair-wise treatment comparisons. Guyatt et al. had reason to believe that a difference of 0.5 on the quality of life scale had clinical relevance. They were then able, for example, to calculate the proportion of patients with an observed difference of at least 0.5 for salmeterol compared to salbutamol and the proportion with an observed difference of 0.5 in favor of salbutamol. The difference between these two proportions is then interpreted as being the net proportion of patients showing a clinically relevant benefit under salmeterol and salbutamol. This figure is then converted into a "number needed to treat" to obtain one clinically relevant benefit.

It is not central to the argument here but attention is drawn to the fact that the number needed to treat is an entirely unsatisfactory way to summarize the results from any clinical trial or meta-analysis. A thorough demolition of this fashionable but inadequate measure is given by Hutton (18). See also Smeeth et al. (19) This particular aspect of the problem will not be considered here. Instead, we concentrate on other difficulties with this general approach.

The mistake here is to interpret this observed difference as saying something about the true effect of the drug at an individual level (20). Suppose that within-patient variability is large, and that differences of 0.5

are common from period to period even when the same drug is given. Then when we observe a difference of 0.5 when comparing two drugs we do not know whether this is a chance difference or a genuine difference. The reason we run a large clinical trial with many patients is that we need that number to tell whether the treatment works at all. It is thus not possible to tell at the individual level whether the treatment works for a given patient. A good discussion of the general difficulty in attempting to do this is given by Cox et al. (21). As we showed in Table 2, for the ordinary cross-over the effects of within-patient variation and patient-by-treatment interaction are not separable.

Such a separation could have been made in our third example. This was a double blind randomized comparison of paracetamol 1 g b.i.d. and diclofenac 50 mg b.i.d. in osteoarthritis reported by March et al. (22). Twenty-five patients, of whom 20 completed the study, were treated in six two-week periods: three times with paracetamol and three times with diclofenac. This study was described as consisting of a series of "n-of-1" trials. However, it can equally be regarded as being a repeated period cross-over. It thus has the structure that would permit separation of patient-by-treatment interaction and within-patient error.

Unfortunately, however, the authors did not carry out a random-effects analysis of these data, preferring instead to perform individual significance tests for each patient. This is a particularly poor approach to analyzing such data (23,24). The authors then concluded that because a clear advantage could only be demonstrated for diclofenac for five patients there were many for whom paracetamol would be adequate.

There are several errors with this approach (22). First, there is a presumption that failure to find a significant difference constitutes a proof of equivalence. Second, the data are not analyzed as a whole, which is inefficient. Third, differences in "significance" are taken as being differences in effect. However, suppose that the effect is identical for every patient and that the power is 25%. Then we would expect that 5 out of every 20 patients would show a significant benefit for diclofenac and 15 would not, despite no difference between patients. Fourth, the generalization of these results from trial to target population is dangerous, especially when the trial has such a narrow basis.

In fact, it is possible to attempt a partial reconstruction of the original data from March et al. and perform a random effects analysis (22). This *does* suggest *some* heterogeneity (if less than the authors' analysis implies). The point, however, is that even with a design that can resolve components of variation effectively we can ascribe too much of the variation to patient-by-treatment interaction if we do not take care in analysis.

A REPEATED PERIODS CROSS-OVER

We now consider our fourth example, an excellent paper in this journal by Shumaker and Metzler (25). Shumaker and Metzler describe a bioequivalence study in which two formulations, a test (T) and a reference (R) of phenytoin, were compared by giving them to 26 healthy volunteers. The four area under the concentration time curve (AUC) values for each subject are plotted in Figure 4. A repeated periods cross-over was used and each subject was allocated to one of two sequences: either RTTR or TRRT, 13 subjects being allocated to each. As Shumaker and Metzler point out, this could be regarded as a trial in which (in a sense) the standard two-period cross-over trial has been replicated twice so that we have sequences RT and TR for the first replicate and TR and RT for the second. This means that for each patient two estimates of the treatment contrasts R versus T can be made: one based on periods one and two and one based on periods three and four. From the point of view of a randomization purist, if this form of analysis is being used, it would probably be better to have used four sequences: the two used and RTRT and TRTR in addition. However, this is a minor criticism. The trial gives a great deal of useful information, as will be shown below.

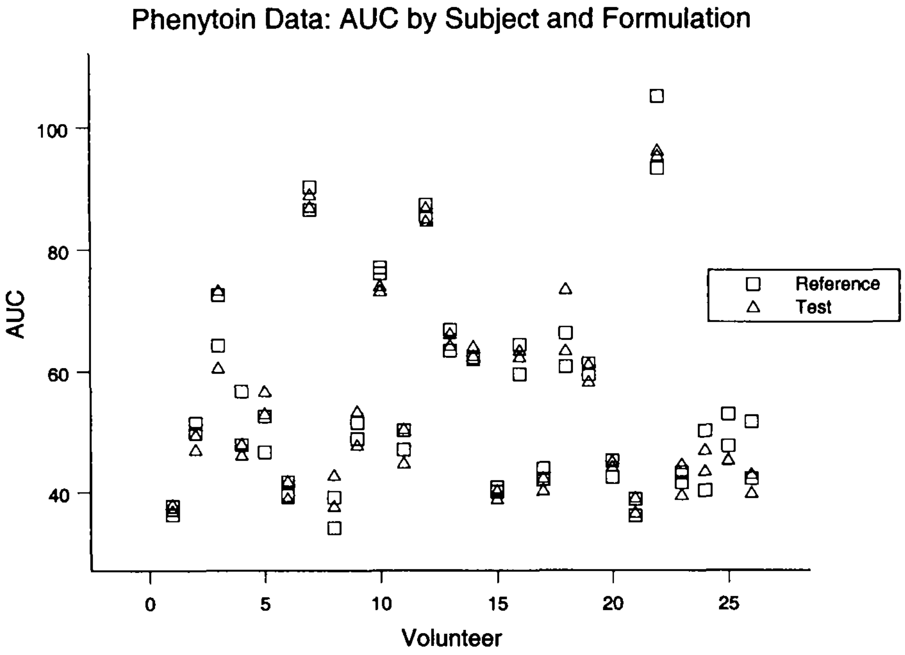

FIGURE 4. AUC for both formulations (reference and test) and both replications for the 26 volunteers in the phenytoin bioequivalence trial.

As is standard for such trials, the principle outcome measure is area under the AUC. If, as a simple way for getting a feel for the data, we calculate the ratio (test/reference) of the AUCs for each replicate then the results are as summarized in Table 4.

The mean relative bioavailability for the two formulations on either replication is very close to one and indeed, since the standard error of the mean ratio is small, a formal analysis of log-transformed ratio shows that the 90% confidence limits (anti-logged) lie well within the conventional limits of equivalence of 80% to 125%. (An alternative is to analyze the original AUCs using the method of Kieser and Hauschke [26]. However, it is the opinion of this author that log-transformation will usually be suitable.) The distribution also appears very stable from one replication to another with minor changes, at most, in any of the statistics.

However, suppose we were to use the method of Guyatt et al. (17). They applied their approach to an ordinary cross-over trial without replication. Thus, if this method is adequate in general, we can apply it to the data from replication one alone. We could then argue that although the mean relative bioavailability is close to 1, there are some subjects for whom it is not so close. Suppose we take as a standard of no practical difference that the ratio should be within the limits 0.95 to 1.053. It then turns out that there are

TABLE 4
Summary Statistics for Relative Bioavailability for the Two Replications

	Replication	
Statistic	**One**	**Two**
Number of observations	26	26
Arithmetic mean	0.999	0.987
Geometric mean	0.996	0.984
Median	0.995	0.993
Minimum	0.813	0.831
Maximum	1.253	1.209
Lower quartile	0.950	0.942
Upper quartile	1.012	1.028
Standard deviation	0.084	0.081
Standard error of mean ratio	0.016	0.016

seven subjects with ratios less than 0.95 and four with ratios in excess of 1.053. We might then be tempted to conclude, as the method of Guyatt et al. would invite us to do, that there are some subjects for whom the test formulation is more bioavailable than the reference and some for whom it is less bioavailable.

However, one very important statistic is missing from the above and that is the correlation coefficient over the 26 subjects and between replications of the relative bioavailability. This correlation coefficient is –0.18! The fact that the correlation coefficient is negative is no doubt a fluke. However, what it strongly suggests is that the observed differences between subjects in replication one has nothing to do with any permanent feature of the subject. It is a purely transient effect, quite possibly largely related to measurement error.

Now suppose that we give a physician the following choice for predicting the relative bioavailability of the two formulations in the second replication for a given subject. She or he can either use the relative bioavailability for replication one for that subject or the global average over all subjects. Which will prove better? The answer is that the better bet is to use the global average and ignore the individual information. This is illustrated in Figure 5, which plots the absolute prediction error for one approach compared to the other. The subjects are numbered. For those who lie to the left of the diagonal line, the prediction error is greater using individual rather than global values, for those who lie to the right the reverse is the case. There are 19 such subjects for whom the global approach is better. Only for seven does the individual value do better. All in all, the evidence is fairly convincing that the global approach is the better one. Of course, since it is impossible to identify for which subjects the individual approach would prove a better prediction (since hindsight cannot be used for prediction!) the best strategy is always to use the global approach.

The trial reported by Shumaker and Metzler (25) thus illustrates graphically (literally) the logical flaw in the method of Guyatt et al. (17). Patient-by-treatment interaction is not identifiable in conventional clinical trials (not even in conventional cross-overs). It may be argued that this case is rather special in that for a bioequivalence trial we do not really expect any difference between treatments. It is thus not surprising that there is no difference at the individual level. The point is, however, that nothing in the structure of the data from the trial if these data are limited to the first replicate alone allows us to identify patient-by-treatment interaction. It is, of course, perfectly possible that such interaction may be important, even in the trial discussed by Guyatt et al. (although it is the opinion of this author that pure within-patient error is likely to be important in that trial). However, nothing in the data forces this conclusion. Attempting to interpret such trials in this way is merely squaring the circle.

THE ROLE OF REPEATED PERIODS CROSS-OVERS

The Shumaker and Metzler study is a valuable example (25). This is not because the topic of individual bioequivalence is important. On the contrary, of almost all areas in drug development in which it would be interesting to study subject-by-treatment interaction this is the least interesting, not least because if the main effect of treatment is small it is implausible that interactive effects involving it could be large (27). However, if individual bioequivalence has been a dead-end as regards practical utility it has not been so as regards theoretical interest. The repeated measures cross-over that has been proposed by the proponents of individual bioequivalence would be the appropriate medium by which to study subject-by-formulation interaction were this matter of interest. Exactly the same design is one of the most powerful means of examining patient-by-treatment interaction in therapeutic trials.

As already discussed, patient-by-treatment interaction provides an upper bound for gene-by-treatment interaction. For example,

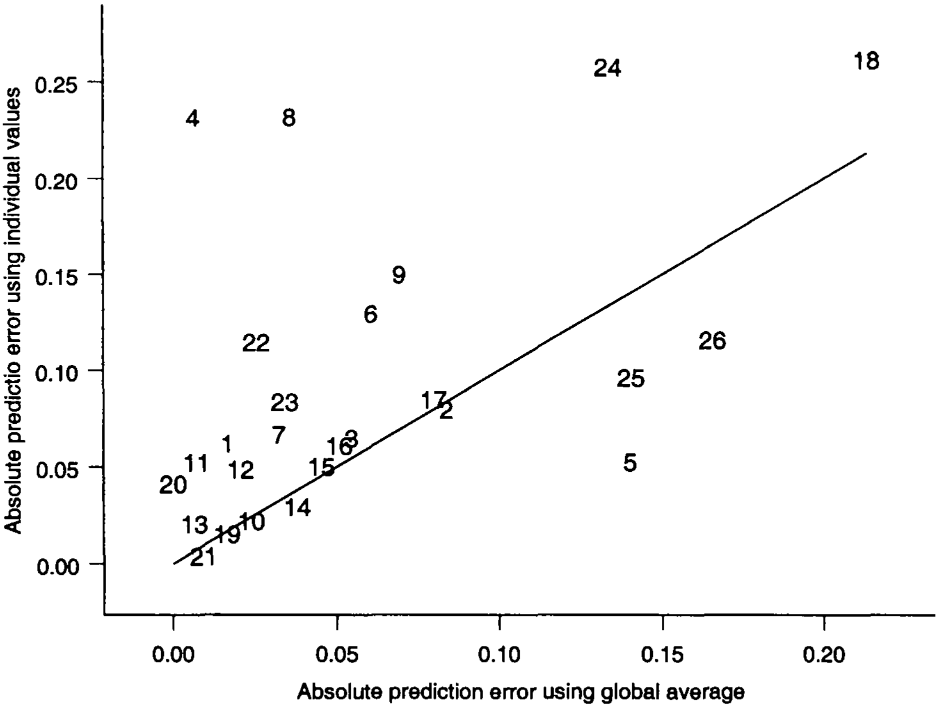

Figure 5. Absolute prediction error for relative bioavailability for the second replication labeled by subject number. The error using the individual results from replication one is plotted against that using the global average.

to take a simple case, it is now known that grapefruit juice interferes with the elimination of many pharmaceuticals (28). It will also be the case that consumption of grapefruit juice will vary from individual to individual (as well, of course, as within individuals!) Unless, therefore, consumption of grapefruit juice is entirely genetically determined, there is some component of the response to such drugs that is modified by an "environmental" factor that varies among individuals and is not genetic. This suggests a possible approach for screening drugs to see whether it is worth investigating genetic components of response, that is, that repeated periods cross-overs should be used to establish the size of the patient-by-treatment interaction for a given pharmaceutical. If this is not large it is not worth undertaking further research to establish the size of gene-by-treatment interaction. This approach has been proposed by Kalow, Ozdemir, Endrenyi, Tang, and co-workers (29,30,31).

Of course, not all indications will be suitable for such an approach. The same limitations will apply that apply to cross-over trials generally (32,33,34), that is to say, the disease must be chronic, the treatments must be reversible, and for practical reasons the necessary period of study must not be so long as to make it impractical from the point of view of the patients who will (ideally) have to have four periods of treatment. Provided suitable precautions are taken, the risk of substantial carry-over can be kept to a minimum. In any case, such trials will not be used for confirmatory proof of efficacy but as part of a general screening strategy. Carry-over would be extremely unlikely to cause one to declare that there was an appreciable interaction where there was not. In the presence of carry-over there might be some slight risk of missing an important interaction but in view of the potential savings this method might bring this risk is acceptable. Suitable designs might be limited to two sequences, ABBA/BAAB, as in the case of Shumaker and Metzler's design (25) but it could also be advantageous to use four sequence designs by adding the sequences ABAB/BABA or even six sequence designs using the sequences AABB/BBAA as well.

A FURTHER DESIGN ISSUE

It is perhaps worth noting at this point that the two aspects of gene-by-treatment interaction alluded to in the opening passage by Sir Richard tend to pull in different directions. The possibility of eliminating a source of variability in order to have "smaller, more effective clinical trials with corresponding cost savings" suggests, if anything, a narrow genetic focus for our trials or at least it suggests that selection of a suitable genetically homogenous group would be at least as efficient, if not more efficient, than stratifying by genotype.

If, on the other hand, we wish to explore fully the extent of gene-by-treatment interaction it may be necessary not only to recruit patients who represent the various genotypes in the target population but also to over-sample the rarer ones. If one wishes to explore the full extent of gene-by-treatment interaction it is necessary to achieve "genetic distance" among the subjects studied and this requires having adequate numbers of the various subgroups. This raises an analogous logistic problem to that which has been noted in connection with attempts to represent women and demographic subgroups "adequately" in clinical trials. However laudable the motive behind this desire, it can lead to a growth in recruitment time that is not practical and may ultimately act against the interest of patients (16,35). Suffice it so say that if the second of Sir Richard's hopes is to be realized for all patients ("individual patients will be targeted with specific treatment and personalized dosing regimens to maximize efficacy and minimize pharmacokinetic problems"), trials will have to get bigger and not smaller.

A RELATED PROBLEM WITH CERTAIN TREATMENT MEASURES

Recently, various statisticians have proposed using treatment measures that describe the extent to which results from one treatment group overlap those of another. These measures have generally been discussed in the context of parallel group trials. For example, Rom and Hwang have defined a measure that is the proportion of similar responses in the two groups (36,37). Hauck and Anderson have suggested using the probability that the outcome from a patient chosen at random from the experimental group is higher than the outcome from a patient chosen at random from the control group (38). A similar suggestion has been made by Chen and Kianifard (39). A much earlier (and very famous) paper by Glass introduced the *effect size*, the ratio of the treatment difference to the within group standard deviation, which he pointed out could be translated into the probability that the observed response in the treatment group will be higher than the mean response in the control group (40).

The danger with these measures comes with their potential interpretation. Consider the measure proposed by Hauck et al. (38), which has been referred to elsewhere as the *individual exceedence probability* (IEP) (37). It is the probability that a measure in the experimental group is *observed* to be higher than the measure in the control group. This must not be confused with the probability that a patient would show, on average, a greater benefit under the experimental treatment than under the control treatment, what we might refer to as the *average benefit probability* (ABP). This would be to make a similar mistake to that made by Guyatt et al. (17) and discussed above. Provided that the individual exceedence probability is greater than 50% the ABP can be as high as 100%. This can be seen very simply by supposing that we take a series of measurements X drawn from a Normal distribution with mean

μ and variance σ^2 and add to each a value τ to form a new series of measurements *Y*, which can then be regarded as having been drawn from a Normal with mean $\mu + \tau$ and variance σ^2. Now imagine that we observe only half the X values at random and wherever we do not observe the X values we observe the Y values instead. This is one possible simple statistical model for what happens in a randomized clinical trial and also corresponds to the Rubin causal model (41,42). Unless τ is large compared to σ we shall observe a considerable overlap of Y and X values but, in fact, for every observed Y, there is an unobserved X that is smaller by τ and for every observed X there is an unobserved Y that is larger by τ. Hence, in general, the degree of overlap does not permit identification of the probability of response.

Even if we are completely chaste in our inferences, are not seduced by the temptation to interpret the IEP causally and continue to interpret it in terms of observations only, there is a further difficulty. We need to identify the population to which the IEP will apply. This will not generally be the target population for the drug. This is because the patients in a clinical trial cannot be regarded as being a random or even representative sample of the target population (16). That being so, the expected value of the sample variance of observations s^2_s will not be the same as that of the variance in the target population, σ^2_p. Even if the treatment effect were completely additive, which is to say, constant from one patient to another, the expected value of Glass's effect size would not be constant from one trial to another unless the variability in these trials was the same (37). But one, say, might be a trial in moderate hypertensives only and another in moderate and severe hypertensives. The effect size would be expected to be less in the latter trial, other things being equal, than in the former, simply because the standard deviation would be expected to be larger.

The only conclusion would then be that the IEP has to be taken to apply to the sample of patients in the trial as actually run over all randomizations. This considerably limits its utility because it makes its applicability extremely local. However, this is not even the end of the difficulties with this measure. Consider, for example, a clinical trial in hypertension with diastolic blood pressure as the target variable which three statisticians analyze. One uses raw outcomes, another uses change scores, and a third (most efficiently) uses the covariance-adjusted outcome. Provided that they stick with the original scale of measurement there is no difference in the expected value for the treatment effect from any of these three analyses although the treatment estimates will be made with different precision. However, ratios of effects to standard deviations can be quite different

CONCLUSIONS

It is not denied that genetic variation in response exists. It is known, for example, that genetic variation can have an important effect pharmacokinetically and that this has implications for adverse reactions (43). Even in terms of pharmacodynamic response it can be highly relevant. For example, it has been claimed that there is major genetic variation among Nigerians as regards angiotensinogen (44). However, it does not follow that all variation that we see is genetic and it is the case that we have rarely designed trials that would separate patient-by-treatment interaction from other sources. This is a situation that could be rectified.

In summary the following conclusions are offered:

1. The temptation to attempt to identify patient-by-treatment interaction when the structure of the clinical trial does not permit it should be resisted,
2. Various methods that have been proposed in the medical literature for doing so are invalid,
3. Although it is possible that pharmacogenomics will enable us to run "cleaner" clinical trials with less variability, we do not know that this will generally be so.

The reason is that for all we know much of the variability we currently see in many clinical trials may not be genetic in origin,

4. Furthermore, the ambition to individualize therapy tends in the other direction. To be realized it may actually require bigger and more difficult trials in which genetic minorities have to be over-sampled, with attendant recruitment problems,
5. We should be very cautious in employing treatment effect measures such as the IEP,
6. A potentially useful and under-exploited design is that of the repeated period cross-over. (Equivalently this can be regarded as a series of n-of-1 trials [23]). This has considerable potential to identify patient-by-treatment interaction and this may help prevent disappointment when hunting for genetic effects (29), and
7. More use should be made of random effect models in analyzing trials and using their results.

Acknowledgment—The author thanks an anonymous referee for helpful comments.

REFERENCES

1. Sykes R. *The Pharmaceutical Industry in the New Millenium: Capturing the Scientific Promise.* London: Centre for Medicines Research; 1977, 1–28.
2. Evans WE, Relling MV. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science.* 1999;286:487–491.
3. Conover WJ, Salsburg DS. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to respond to treatment. *Biometrics.* 1988;44:189–196.
4. Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput Biomed Res.* 1972;5:411–459.
5. Hartl DL. *A Primer of Population Genetics.* Sunderland, MA: Sinauer Associates, Inc; 2000.
6. Elston RC. Introduction and overview. *Stat Methods Med Res.* 2000;9:527–541.
7. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation [see comments]. *J Clin Epidemiol.* 1996;49:395–400.
8. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics.* 1985;41:361–372.
9. Senn SJ, Harrell F. On wisdom after the event [comment] [see comments]. *J Clin Epidemiol.* 1997;50: 749–751.
10. Senn SJ, Harrell FE Jr. On subgroups and groping for significance [letter; comment]. *J Clin Epidemiol.* 1998;51:1367–1368.
11. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. On reaching the tunnel at the end of the light [comment] [see comments]. *J Clin Epidemiol.* 1997;50: 753–755.
12. Galbraith RF. Graphical display of estimates having differing standard errors. *Technometrics.* 1988;30: 271–281.
13. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical-trials. *Stat Med.* 1988;7:889–894.
14. Galbraith RF. Some applications of radial plots. *J Am Stat Assoc.* 1994;89:1232–1242.
15. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Stat Med.* 1996; 15:619–629.
16. Senn SJ. *Statistical Issues in Drug Development.* Chichester: John Wiley; 1997.
17. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials [see comments]. *Br Med J.* 1998;316: 690–693.
18. Hutton JL. Numbers needed to treat: properties and problems (with comments). *J Roy Stat Society A.* 2000;163:403–419.
19. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses—sometimes informative, usually misleading [see comments]. *Br Med J.* 1999;318:1548–1551.
20. Senn SJ. Applying results of randomised trials to patients. N of 1 trials are needed [letter; comment]. *Br Med J.* 1998;317:537–538.
21. Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Spiegelhalter DJ, Jones DR. Quality-of-life assessment—Can we keep it simple. *J R Stat Soc Ser A-Stat Soc.* 1992;155:353–393.
22. March L, Irwig L, Schwarz J, Simpson J, Chock C, Brooks P. n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. *Br Med J.* 1994;309:1041–1045; discussion 1045–1046.
23. Senn SJ. Suspended judgment n-of-1 trials. *Control Clin Trials.* 1993;14:1–5.
24. Senn SJ, Bakshi R, Ezzet N. n of 1 trials in osteoarthritis. Caution in interpretation needed [letter; comment]. *Br Med J.* 1995;310:667.
25. Shumaker RC, Metzler, CM. The phenytoin trial is a case study of "individual bioequivalence." *Drug Inf J.* 1998;32:1063–1072.
26. Kieser M, Hauschke D. Statistical method for demonstrating equivalence in crossover trials based on the ratio of two location parameters. *Drug Inf J.* 2000;34:563–568.
27. Senn SJ. In the blood: proposed new requirements for registering generic drugs. *Lancet.* 1998;352:85–86.

28. Kane GC, Lipsky JJ. Drug-grapefruit juice interactions. *Mayo Clin Proc.* 2000;75:933–942.
29. Kalow W, Tang BK, Endrenyi L. Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics.* 1998;8:283–289.
30. Kalow W, Ozdemir V, Tang BK, Tothfalusi L, Endrenyi L. The science of pharmacological variability: an essay. *Clin Pharmacol Ther.* 1999;66:445–447.
31. Ozdemir V, Kalowa W, Tang BK, Paterson AD, Walker SE, Endrenyi L, Kashuba AD. Evaluation of the genetic component of variability in CYP3A4 activity: a repeated drug administration method. *Pharmacogenetics.* 2000;10:373–388.
32. Senn SJ. Crossover Design. In: *Encyclopedia of Biopharmaceutical Statistics.* SC Chow and JP Liu, eds. New York: Marcel Dekker; 2000, 142–149.
33. Senn SJ. Cross-over trials. In: *Encyclopedia in Biostatistics.*, P Armitage and T Colton, eds. New York: Wiley; 1998, 1033–1049.
34. Senn SJ. *Cross-over Trials in Clinical Research.* Chichester: John Wiley; 1993.
35. Meinert CL, Gilpin AK, Unalp A, Dawson C. Gender representation in trials. *Control Clin Trials.* 2000;21:462–475.
36. Rom DM, Hwang E. Testing for individual and population equivalence based on the proportion of similar responses [see comments]. *Stat Med.* 1996;15:1489–1505.
37. Senn SJ. Testing for individual and population equivalence based on the proportion of similar responses [letter; comment]. *Stat Med.* 1997;16:1303–1306.
38. Hauck WW, Hyslop T, Anderson S. Generalized treatment effects for clinical trials. *Stat Med.* 2000; 19:887–899.
39. Chen M, Kianifard F. A nonparametric procedure associated with a clinically meaningful efficacy measure. *Biostatistics.* 2000;1:293–298.
40. Glass, GE. Primary, secondary and meta-analysis of research. *Educat Research.* 1976;5:3–8.
41. Rosenbaum PR, Rubin DR. The central role of the propensity score in observational studies for causal effect. *Biometrika.* 1983;70:41–55.
42. Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. *J Educat Psychol.* 1974;66:688–701.
43. Meyer UA. Pharmacogenetics and adverse drug reactions. *Lancet.* 2000;356.
44. Guo X, Rotimi C, Cooper R, Luke A, Elston RC, Ogunbiyi O, Ward R. Evidence of a major gene effect for angiotensinogen among Nigerians. *Ann Hum Genet.* 1999;63:293–300.