Do commonly used clinical trial designs reflect clinical reality?

Elihu Estey

Fred Hutchinson Cancer Research Center and University of Washington School of Medicine, Seattle, USA

ABSTRACT

This paper contends that commonly used clinical trial designs do not reflect clinical reality as viewed by patients or physicians. Specifically, randomized phase III designs focus on improvements that are more significant statistically than medically and put an emphasis on avoiding a false positive result that is more appropriate for diseases that are curable, in contrast to acute leukemias. The resultant large sample sizes needed for each treatment restrict the trial to one or two new treatments, although historical reality suggests the difficulty in knowing, without clinical data, whether these are the best of several new treatments. The *p* value-based statistics discourage use of data from previous patients in the trial to inform treatment of subsequent patients, contravening patients' assumptions. Standard phase II trials focus on a single outcome, ignoring the complexity of medical practice, and ignore prognostic heterogeneity. Finally, although patients are more interested in whether a new treatment is better than another, rather than whether it is active, randomization between different treatments does not begin until phase II trials have been completed. This paper proposes alternatives based on the Bayesian statistical approach. The thesis that I will develop here is that commonly used clinical trial designs are unrealistic in the sense that they do not correspond well to patients' views of medical practice and greatly over-simplify such practice. By emphasizing Bayesian rather than *p* value-based statistics and focusing on acute myeloid leukemia, I hope to familiarize physicians with some of the many new published designs that address these problems.

Key words: *MLL*, proteins, leukemia.

Citation: Estey E. Do commonly used clinical trial designs reflect clinical reality? Haematologica 2009.94:1435-1439. doi:10.3324/haematol.2009.011411

©2009 Ferrata Storti Foundation. This is an open-access paper.

©Ferrata Storti Foundation

The standard phase III trial

These trials typically randomize approximately 400 patients between two therapies.[1-3] This relatively large number is required to detect relatively small improvements with a false positive rate less than 5% (p<0.05) and a false negative rate less than 20% (80% power). For example, the trials in references 1-3 targeted increases in median event-free survival (EFS) or survival of 6-12 months, in 2-year EFS or survival of from 10% to 20% and in complete remission (CR) rate from 50% to 65%. Consider the relevance of a 6-month improvement in survival to an otherwise healthy 65-year old man with untreated acute myeloid leukemia (AML). Such a patient might expect to live another 15 years if he did not have AML but only another one half-year if he is randomized to a standard treatment arm. In such a case, he only retains 0.5/15 (3%) of his normally remaining life expectancy. If he is randomized to the investigational arm and it is *successful*, he gains another half-year and now retains 1/15 (7%) of his life expectancy.

While statistically significant, I doubt many patients would consider this result medically significant. Hence, the targeted improvement does not reflect clinical reality. The choice of a false positive rate of 0.05 but a false negative rate of 0.20 implies a preference for more protection against a false positive than a false negative result. This is quite sensible when satisfactory treatment exists for the disease in question, and hence, replacement of this standard with a falsely positive new therapy is particularly undesirable. However, because there is no satisfactory treatment for most patients with AML, the medical risk of a false positive is much less. Indeed, the near universal choice of p=0.05 and power=80%, regardless of the disease in question, ignores the reality that diseases vary considerably in curability.

Consequently, phase III AML trials should perhaps seek more clinically meaningful improvements and permit higher p values. Although this formulation would result in loss of power to detect relatively small advances, I question whether leukemia therapeutics advances in such small increments. In particular, it would appear that quantum therapeutic advances

Manuscript received on May 14, 2009. Manuscript accepted on May 21, 2009.

Correspondence: Elihu Estey, Fred Hutchinson Cancer Research Center and University of Washington School of Medicine, Seattle, USA. E-mail: eestey@u.washington.edu

are not infrequent, as with all-trans retinoic acid (ATRA) and arsenic trioxide (ATO) for APL, 2-chlorodeoxyadenosine for hairy cell leukemia, high-dose ara-C and likely gemtuzumab ozogamycin for CBF AML, and imatinib for chronic myeloid leukemia (CML). Even if there were value in retaining sufficient power to detect small advances, the added value may not justify the necessary sample sizes, which prevent expeditious completion of trials and simultaneous investigation of a large number of new therapies.

P value-based versus Bayesian approaches

Patients naturally prefer *adaptive* designs, those that permit treatment decisions for subsequent patients in a trial to be based on results in previous patients. However, *p* value-based designs tend to discourage frequent examination of incoming data. This reflects the inextricable link between *p* value and trial design, such that the same data can produce different *p* values depending on the particular design used (Table 1).[4-7] For example, it is well-known that the probability of finding an association at $p<0.05$ increases purely by chance as the number of tests of significance that are performed increases.[8,9]

Accordingly, interim analyses of clinical trials are generally performed at *p* values much less than 0.05 in order to preserve an approximately 0.05 level of significance at the final analysis. For example, the design proposed by Fleming et al. stops a trial, declaring one arm superior, only with *p* values of 0.005, 0.006, 0.007, and 0.009 at the 1st, 2nd, 3rd, and 4th of 4 interim analyses, respectively. This of course makes it difficult to stop 1:1 randomization to an arm, even when the probability that that arm is inferior is greater than 90%, leading most patients to prefer randomization to the better arm.

The dependence of *p* value on trial design is such that, in a case in which the final planned analysis yields a *p* value of 0.051, but in which subsequently obtained data strengthen the evidence in favor of a difference, these data cannot be used since they were not obtained as part of the planned experiment.

Table 1. Definitions.

p value: the probability of observing the data or more extreme data under the null hypothesis; the latter states that there is truly no difference between two treatments. For example, assume the null hypothesis is that, among 10 people, 5 would prefer red and 5 blue. In fact, 8 preferred red and 2 blue. The *p* value is calculated as the probability that, under the null hypothesis, 8/10 would prefer red + the probability that 9/10 would prefer red + the probability that all 10 would prefer red.

Bayes's theorem states:

$$P(B|A) = \frac{[P(A|B)]\ [P(B)]}{[P(A|B)]\ [P(B)] + [P(A|\text{not } B)]\ [P(\text{not } B)]}$$

where P is probability, B is a hypothesis, and A are observed data. P(B) is known as the prior probability of the hypothesis, while P(B | A) is the posterior probability of the hypothesis. Thus, the *p* value is based on the probability of data given a hypothesis while Bayesians compute the posterior probability of a hypothesis given data. Physicians often mistakenly believe that a *p* value is a Bayesian posterior probability.

The Bayesian approach provides flexibility, and in particular, encourages interim analyses. The approach begins with parameters, such as the probability of CR or, when comparing two treatments, the probability that the relative risk of survival is greater than 1.0. These parameters (denoted here by θ) are random quantities, with probability distributions describing one's uncertainty about them. One begins with a prior distribution, p(θ), that characterizes the uncertainty about θ before observing any data. The second Bayesian quantity is the likelihood, L(data | θ), which describes the probability of observing any specified data given any value of θ; examples of likelihoods are the binomial distribution for binary events and the normal (bell-shaped) distribution for continuous variables. Bayes's theorem multiplies the prior by the likelihood of observing the data given the parameter to arrive at a *posterior* distribution of θ, which describes uncertainty about θ after observing the data (Figure 1). In contrast to *p* value-based methods, Bayesian inference is not affected by the experimental design since data only enter inferences through the likelihood function. Consequently, when making decisions or inferences based on accruing data, Bayes's theorem may be repeatedly applied, with the posterior at each stage becoming the prior for the next stage. The proba-

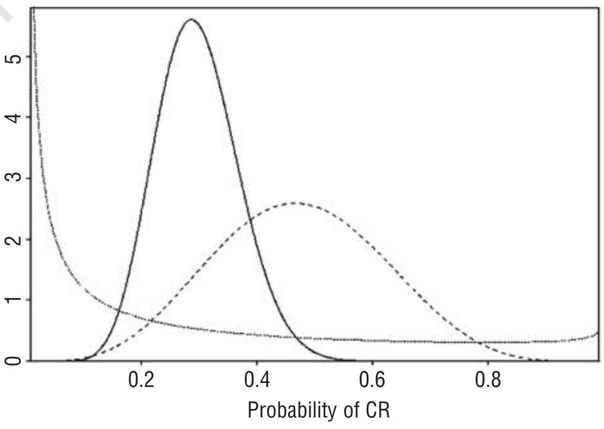

Figure 1. Bayesian probability distributions using a trial of a new therapy in relapsed acute myeloid leukemia as an example. The values on the horizontal axis are different probabilities of complete remission. The values on the vertical axis represent the weight assigned to each CR probability. Prior to treatment, although the average CR rate is thought to be 20%, some credence is assigned to each probability of CR (prior probability distribution, dotted line). After observing 5/10 CRS (first posterior probability distribution, dashed line), the average CR rate is close to 50% and no credence is given to CR rates less than 10% or greater than 90%, reflecting the impact of the observed data on the prior. Thus, the posteriors become successively more informative as the data accumulate, and shift to reflect the overall average behavior of the data. After observing 7 CRs in the next 30 patients (total 12 CRs in 40 patients), the average CR rate is approximately 30% and no credence is given to a CR rate greater than 60% (2nd posterior probability distribution, solid line). Computing the proportion of the area under the curve that is to the right of a CR rate of 0.4 gives the current probability that the CR rate is greater than 0.4. This probability can be used to make treatment decisions.

©Ferrata Storti Foundation

bility distributions in this sequence become increasingly informative about θ as the data accumulate. This process, known as *Bayesian learning* (Figure 1), is especially useful in sequential data monitoring during a clinical trial. The current posterior probability distribution may be used to modify doses, unbalance a randomization in favor of a treatment with relatively superior performance, or terminate a trial early due to either superiority of a treatment or futility. The Bayesian approach's flexibility can be appreciated by contrasting its ability to incorporate data obtained subsequent to trial completion with the *p* value approach's inability to do this, as noted above.

A significant issue with the Bayesian approach is setting prior probabilities. In Figure 1, we made the prior non-informative reflecting a lack of any information about response to the new drug. However, it might be contended that the prior should be more informative, incorporating knowledge of previous trials with other drugs in relapsed AML. The choice of prior obviously influences computation of the posterior – the more informative the prior, the more data needed to influence the posterior. The designs described below generally use non-informative priors. A more detailed presentation would describe how selection of different priors influences the posterior.

Adaptive randomization

Bayesian designs for adaptive randomization repeatedly use interim data to compute the probability that one arm of a randomized trial is better than the other(s), unbalancing the randomization to favor the likely better treatment.[10,11] If this probability crosses a pre-specified boundary, the inferior arm is shut down before the maximal sample size is reached. However, it may reopen if further analyses indicate that results with the open arm(s) are deteriorating such that the probability that this arm(s) is superior has decreased.

A trial adaptively randomizing patients over age 50 with untreated AML among idarubicin + ara-C (IA, the standard), troxacitabine + ara-C (TA), and troxacitabine + idarubicin (TI) illustrates the process.[12] The first 15 patients were randomized fairly among the three arms. As each patient after the 15th entered the trial, we computed the posterior probability that the CR rate with IA was greater than or equal to 10% better than that with TA or TI. If this probability was less than 0.15, accrual to IA was suspended. If in contrast the posterior probability was greater than 0.85 that the CR rate with TA or TI was greater than or equal to 10% worse with IA, accrual to either TA or TI was suspended. Depending on results in arms that remained open, a closed arm could re-open. A maximum of 75 patients were to be randomized. The TI arm closed and remained closed after the first 5 patients failed to respond, while the TA arm closed and remained closed after the CR rate was 3/11, at which time the CR rate in the IA arm was 10/18. If the 34 patients who had been entered on the trial when both TA and TI arms were closed had been randomized fairly, 11 patients would have received each of TA, TI, and IA. With adaptive randomization, only 16, rather than 22 patients, received the inferior TA or TI arms, probably corresponding with how patients visualize clinical practice.

The possibility certainly exists that stopping arms so early might lead to a false negative conclusion. Beginning adaptive randomization only once 15 to 20 patients have been randomized equally among the various treatment arms reduces the problem. At any rate, it is critical to examine how the design performs under various clinical scenarios, that is, what are its operating characteristics (OC). OC include the probabilities that the design will correctly select a truly superior treatment or incorrectly select a truly inferior treatment, as well as the median number of patients treated on each arm. If clinicians feel the OC are unsatisfactory, the parameters above, such as the criterion probabilities of 0.15 or 0.85, or the number of patients to be fairly randomized are changed until desirable OC are obtained. Table 2 illustrates 1,000 computer simulations for two scenarios in the IA versus TA versus TI trial. In the first, the true CR rates with TA, IA, and TI are 50%, 40%, and 30%, respectively; hence, the correct conclusion is that TA is superior. As parameterized above, the probability was 80% that the design would reach the correct conclusion, corresponding to a power of 80%. In contrast, if the true CR rates were 30%, 40%, and 30% with TA, IA, and TI, respectively, the probability that the design would correctly select IA as superior was only 10%. Hence, in this case, the design provided much more protection against a false negative than a false positive. The false positive rate could have been decreased by eliminating the requirement that, with high probability, TA or TI be at least 10% worse than IA before either of these arms would close. However, this would have also increased the false negative rate contrary to the desire of the clinical investigators to maintain a low false negative rate.

As outlined above, adaptive randomization fails to account for the possible imbalance in prognostic covariates between patients randomized on each arm. This issue has recently been addressed, together with how adaptive randomization may be used with censored data as might arise when survival is the endpoint.[13] In any event, implementation of adaptive randomization requires that patients only infrequently present for randomization before there has been sufficient opportunity to observe the outcome in previous patients.

Accounting for prognostic heterogeneity in single arm trials

New drugs are typically tested in single-arm phase II trials before investigation in phase III. The most commonly used design for single arm phase II trials is the Simon 2-stage (S2S) design.[14-15] Rates of no interest (known as p0), typically corresponding to the historical

Table 2. Operating characteristics for IA *vs.* TA *vs.* TI trial.

True CR rates			Correct conclusion	Probability correct conclusion
TA	IA	TI		
50%	40%	30%	TA superior	80%
30%	40%	30%	IA superior	10%

©Ferrata Storti Foundation

rate, and of interest (p1) typically 0.15 to 0.20 higher than p0 are specified, together with maximum false positive and false negative rates (typically 0.10). These parameters determine the number of patients treated in the first stage and the minimum number of responses needed to proceed to a second stage of specified number. After the latter is completed, a drug is accepted if the number of responses is greater than the specified minimum.

The S2S unrealistically assumes that treated patients have homogeneous prognoses. Certainly, in AML this is unlikely to be the case.[16] Hence reliance on the S2S risks declaring drugs inactive when they might have been found active had a better prognostic group been treated. Conducting separate phase II trials in distinct prognostic groups is time consuming and does not allow information gained in one prognostic group to affect the trial in a second prognostic group.

A method that accounts for treatment-prognostic subgroup interactions has been proposed, specifically using data from the trial to estimate the degree to which the results in the different subgroups can be combined.[17] There are two levels of prior probability distributions (*hierarchical Bayes*). The first is the usual probability of response to a drug in each of, for example, two prognostic groups. The second quantifies prior belief that the response in one prognostic group can inform the probability of response in the other. As usual, these priors are updated as the trial proceeds.

Consider a hypothetical trial of a new drug in relapsed AML. Actual historical data indicated a response rate of 21% in 169 patients. This rate was 11% (118 patients) if initial CR duration was less than one year but 43% (51 patients) if initial CR duration was greater than one year. The goal was to increase response rate to 31% (absolute increase of 0.2) in the worse prognostic group and to 58% (absolute increase of 15%) in the better prognostic group. Since the historical data suggest that 69% of patients will be in the worse group, the overall targeted improvement is [0.20×0.69}+[0.15×0.31]=0.18. Thus, an S2S design would set 21% as p0 and 0.21+0.18=0.39 as p1. Setting the nominal false positive and false negative rates at 0.10, the S2S would treat 22 patients in a first stage, and the trial would stop if less than five responses occurred. If greater than four responses occurred, an additional 21 patients would be enrolled and the drug declared a success if responses were seen in more than 12/43 patients. Thus, to make the proposed design (hereafter, STI because it examines subgroup-treatment interactions) comparable to the STS, we specify that STI will also take its first look after 22 patients have been evaluated and will also set its false negative rate at 0.10.

Table 3 compares the operating characteristics of the STI and S2S designs. In Table 3A, the new drug achieves its goal in the better but not the worse group. Because the S2S does not consider interactions between prognostic subgroups and treatment, it has the same probability (0.75) of rejecting the drug in both groups. In contrast, the STI is less likely to reject the drug in the better group and more likely to reject it in the worse group. Furthermore, 52% of the patients treated with STI will be in the better group versus only 29% with S2S. Table 3B illustrates that, in the case in which the desired improvement occurs in the worse but not the better group, STI is more likely to accept and reject the drug in the appropriate subgroups. Although conducting separate Simon 2-stage designed trials in better and worse subgroups corrects this problem, S2S's inability to allow results in one subgroup to affect the conduct of the trial in the other subgroup continues to result in a smaller proportion of patients belonging to the group where treatment seems more effective relative to historical data.

Monitoring multiple outcomes

The great majority of clinical trials specify one *primary* outcome, such as toxicity, response rate, or survival. Stopping rules are based only on the primary outcome. This formulation appears unrealistic, ignoring the complexities of medical practice and clinical research. For example, because phase I trials are often quite small and, unrealistically, fail to account for covariates other than dose associated with toxicity, knowledge of toxicity is often incomplete after phase I.[18-20] It follows that it is desirable in phase II to formally measure both response and toxicity and allow stopping based on either outcome. Consider also a trial of a new therapy, postulated to be less toxic than standard 3+7, in older patients with untreated AML. While the reduced toxicity might improve survival relative to 3+7, it might also reduce CR rate, with long-term survival most likely in patients achieving CR.[21] However, some decrease in CR rate would be accepted provided survival increased. Thus, the trial would formally monitor both survival and CR, stopping if the decrement in CR rate appeared too great or the increase in survival insufficient. The proportion of eligible patients who actually enrol on a trial is often relatively low due to selection bias. The consequences of such bias might be reduced were trials to stop if it appeared likely that they were only relevant for a small subset of the eligible population. Designs that monitor multiple outcomes are readily available.[22,23]

Testing more new therapies and allowing earlier comparison of these

Patients are more interested in whether one therapy is better than another than whether either therapy is *active*.

Tables 3. Comparative operating characteristics of STI and Simon 2-stage (S2S) designs.

	True	Probability (Reject)		Mean #Pts	
Subgroup	CR Rate	S-TI	S2S	S-TI	S2S
Better	0.58	0.10	0.75	21	10
Worse	0.11	0.90	0.75	19	25

	True	Probability (Reject)		Mean #Pts	
Subgroup	CR Rate	S-TI	S2S	S-TI	S2S
Better	0.43	0.50	0.26	13	11
Worse	0.31	0.10	0.26	27	30

©Ferrata Storti Foundation

Because comparison is best done through randomization, it has been proposed that randomization begin earlier than is now the case. In particular, selection designs have been proposed in which a relatively small number of patients are randomized among several new therapies.[22,24] The rationale is that, although many new therapies are available that may be tested in different schedules and combinations, pre-clinical rationale is an imperfect guide to selecting which new drug to compare with a standard. Thus, a compelling pre-clinical rationale did not exist for arsenic trioxide in APL, fludarabine in CLL, and cladribine in hairy cell leukemia, while many drugs that failed clinically were accompanied by seemingly unassailable rationales. A Bayesian selection design randomizes 45 to 80 patients among three to four therapies. Each therapy begins with the same prior probability distribution. As patients are treated, the priors are updated with these posteriors used to shut down accrual to an arm if, for example, the probability that its true response rate is greater than 20% worse than a competing arm is high. At the end of the trial, the arm with the highest response rate among those not shut down is selected for further study, perhaps in comparison to standard therapy.

Such selection designs are often criticized as *underpowered phase III trials*. Examination of selection designs' operating characteristics indicate that, in a scenario where three drugs have the same true response rate and the fourth provides an absolute 20% improvement, the probability of correctly selecting the fourth drug (that is the probability that it will not stop early plus the probability that it will have the highest response rate at the end of the trial) is only about 60%. This of course contrasts with the aforementioned 80% power typical of randomized trials, involving, for example, a new drug versus a standard. However, the 80% figure is purely nominal, ignoring the process used to select the new drug. Assume that four new therapies were available for comparison with a standard, and that because pre-clinical rationale cannot substitute for clinical data in the selection process, each was equally likely to be useful clinically. It follows that the probability of correctly selecting the best drug was 25%. This 25% is ignored in the computation of 80% power; if it were not, the power of the trial would be 25%×80%=20%.

Thus, the selection design's 60% probability of correct selection should be viewed, not in relation to 80% power, but in relation to the 25% probability of correct selection that it would obtain in the absence of the selection design. Recognizing these issues, the Medical Research Council-sponsored trials in AML in the United Kingdom are employing selection designs rather than more conventional phase III designs.

References

1. Baer M, George S, Dodge R, O'Loughlin KL, Minderman H, Caligiuri MA, et al. Phase 3 study of the multidrug resistance modulator PSC-833 in previously untreated patients 60 years of age and older with acute myeloid leukemia: Cancer and Leukemia Group B Study 9720. Blood 2002;100:1224-32.
2. Rowe J, Neuberg D, Friedenberg W, Bennett JM, Paietta E, Makary AZ, et al. A phase 3 study of three induction regimens and of priming with GM-CSF in older adults with acute myeloid leukemia: a trial by the Eastern Cooperative Oncology Group. Blood 2004;103:479-85.
3. van der Holt B, Lowenberg B, Burnett A, Knauf WU, Shepherd J, Piccaluga PP, et al. The value of the MDR1 reversal agent PSC-833 in addition to daunorubicin and cytarabine in the treatment of elderly patients with previously untreated acute myeloid leukemia, in relation to MDR1 status at diagnosis. Blood 2005;106:2646-54.
4. Berger J, Berry D. Statistical analysis and the illusion of objectivity. Am Sci 1988;76:159-165.
5. Berry DA, Stangl DK, editors. Bayesian Biostatistics. New York: Marcel Dekker; 1996.
6. Goodman SA. Toward Evidence-Based Medical Statistics,1: the p-value fallacy. Ann Intern Med 1999; 130:996-1004.
7. Goodman SA. Toward Evidence-Based Medical Statistics, 2: the Bayes factor. Ann Intern Med 1999;130: 1005-13.
8. Hilsenbeck S, Clark G, McGuire W. Why do so many prognostic factors fail to pan out? Breast Cancer Res Treat 1992;22:197-206.
9. Altman D, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Nat Cancer Inst 1994;86:829-35.
10. Berry D, Eick S. Adaptive assignment vs. balanced randomization in clinical trials: a decision analysis. Stat Med 1995;14:231-46.
11. Thall P, Wathan J. Practical Bayesian adaptive randomization in clinical trials. Eur J Cancer 2007;43:859-66.
12. Giles F, Kantarjian H, Cortes J, Garcia-Manero G, Verstovsek S, Faderl S, et al. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. J Clin Oncol 2003;21:1722-7.
13. Cheung YK, Inoue LY, Wathen JK, Thall PF. Continuous Bayesian adaptive randomization based on event times with covariates. Stat Med 2006;25:55-70.
14. Simon R. Optimal two-stage designs for phase 2 clinical trials. Controlled Clin Trials 1989:10:1-10.
15. Thall PF, Simon R. Incorporating historical control data in planning phase 2 clinical trials. Stat Med 1990;9:215-28.
16. Estey E, Dohner H. Acute myeloid leukaemia. Lancet 2006;368:1894-907.
17. Wathen J, Thall PF, Cook J, Estey E. Accounting for patient heterogeneity in phase 2 clinical trials. Stat Med 2008;27:2802-15.
18. Thall P, Lee S. Practical model-based dose-finding in phase I clinical trials: methods based on toxicity. Int J Gynecol Cancer 2003;13:251-61.
19. Rogatko A, Babb JS, Wang H, Slifker MJ, Hudes GR. Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. Clin Cancer Res 2004;10:4645-51.
20. Thall PF, Nguyen HQ, Estey E. Patient-specific dose finding based on bivariate outcomes and covariates. Biometrics 2008;64:1126-36.
21. Estey E, Garcia-Manero G, Giles F, Cortes J, O'Brien S, Kantarajian H. Clinical relevance of CRp in untreated AML. Blood 2005:106:Abstract 541.
22. Estey E, Thall PF. New designs for phase 2 trials. Blood 2003;102:442-8.
23. Thall PF, Estey E, Sung H. A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. Invest New Drugs 1999:17:155-67.
24. Thall PF, Estey EA. Bayesian strategy for screening cancer treatments prior to phase 2 clinical evaluation. Stat Med 1993; 12:1197-211.

©Ferrata Storti Foundation