Cancer Chemother Pharmacol (1997) 40 (Suppl): S9–S12

ORIGINAL ARTICLE

Elihu Estey · Peter Thall · Cynthia David

Design and analysis of trials of salvage therapy in acute myelogenous leukemia

Abstract Results obtained with a given regimen in relapsed or refractory acute myelogenous leukemia (AML) are variable. This often reflects variability in patient selection. We have developed a system to account for such variability that stratifies patients with refractory or relapsed AML into four groups: group 1, first complete response (CR) duration ≥ 2 years and receiving first salvage treatment (S1); group 2, first CR duration 1–2 years and receiving S1; group 3, first CR duration 0–1 years and receiving S1; and group 4, first CR duration 0–1 years and receiving S2, S3, or S4 after failing S1. CR rates achieved in the four groups are 73%, 47%, 14%, and 0, respectively. This system is useful for comparing results obtained with different therapies and for assigning patients to treatment. At our institution, patients in group 4 are enrolled in phase I studies, and phase I$^1/_2$–II studies are carried out separately in patients in groups 2, 3, and 4. A phase I1/2 study refers to one in which the intent is to select for phase II therapies emerging from phase I trial results. The design is Bayesian, and although false-negative rates are relatively high, they are lower than those obtained if a drug for phase II testing is arbitrarily selected.

Key words Relapsed/refractory acute myelogenous leukemia · Stratification · Bayesian design · Phase I trials · Phase II trials

Work presented at the 12th Bristol-Myers Squibb Nagoya International Cancer Treatment Symposium, "New therapeutic strategies for higher cure rates: High-dose therapy and new therapeutic modalities," 4–5 October 1996, Nagoya, Japan

E. Estey · P. Thall · C. David
University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA

E. Estey (✉)
Department of Hematology, Box 61, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe, Houston, TX 77030, USA
Tel. +1 713- 792-7544; Fax +1 713- 794-4297

Introduction

This paper addresses two common problems in clinical research in acute myelogenous leukemia (AML): first, the variability in results produced by a given therapy, and second, the selection of drugs for phase II studies.

Variability of the results of a given therapy

The problem of variability in results produced by a given therapy is illustrated by trials with 2-chlorodeoxyadenosine, which produced a 47% complete response (CR) rate in 17 children with "relapsed or refractory" AML [6]. On the basis of these results we gave the drug to 21 adults with the same diagnosis; however, no CR was observed [4]. Since the 95% confidence limits for the two response rates do not overlap (23–72% versus 0–17%), the disparate results are not due to random variation but most likely reflect differences in first remission duration between patients in the two studies. These durations were a median of 21 months in the adult study but of only 21 weeks in the pediatric study.

The importance of first CR duration on the outcome of therapy is well known but is often not reported in a consistent fashion. In 1990, Hiddemann et al. [3] defined refractory AML in the hope that use of this definition would facilitate comparisons of different studies. They defined AML as refractory if it failed to respond to initial induction therapy, recurred within 6 months of initial CR, or had recurred more than once. Although certainly an improvement, this definition may nonetheless be insufficiently precise. Clinical trials with the combination of cyclosporine, daunorubicin, and high-dose cytarabine (Ara-C) are illustrative. Using Hiddemann et al.'s definition, List et al. [5] observed a CR rate of 82% [95% confidence interval (CI) 48–98%] in refractory AML. Using the same treatment plan, we saw a CR rate of 0 (95% CI 0–46%) in identically defined refractory AML [4].

To produce more homogeneous groups of patients we analyzed outcome (CR or no CR) in 206 patients (median age 56 years) who received chemotherapy without allogeneic transplantation for relapsed or primary refractory AML, excluding acute promyelocytic leukemia (APL), at M.D. Anderson Cancer Center between 1991 and 1994 [1]. For first salvage, 68% of the patients received conventional (principally high-dose Ara-C-based) regimens; the remaining 32% received investigational regimens [usually single agents such as topotecan, 2-chlorodeoxyadenosine (CDA), CI 973, or Taxol]. The overall CR rate was 23%. Of the 206 patients, 93 received a second salvage attempt either at second relapse or after failing the first attempt. Of these, 43% received conventional regimens and 57% investigational regimens; the overall CR rate was 11%. A total of 40 patients received a third regimen and 17 patients a fourth salvage regimen, resulting in CR rates of 10% and 6%, respectively.

Analysis of these data distinguished four groups: (1) patients with an initial CR duration of >2 years who were receiving their first salvage attempt (15 patients, CR rate 73%, 95% CI 45–92%), (2) patients with an initial CR duration of 1–2 years who were receiving their first salvage attempt (30 patients, CR rate 47%, 95% CI 28–66%), (3) patients with a first CR duration of <1 year or with no initial CR who were receiving their initial salvage attempt (160 patients, CR rate 14%, 95% CI 8–21%), and (4) patients with an initial CR duration of <1 year or with no initial CR who were receiving a second or subsequent (up to and including a fourth) salvage regimen and had not responded to a first salvage attempt (58 patients, 96 salvage attempts, CR rate 0, 95% CI 0–4%). Other patients, e.g., those who responded to a first salvage attempt and underwent a second salvage regimen, were excluded from consideration due to the small number of patients available, although there are suggestions that patients who respond to the first regimen may be relatively responsive to the second (3/7 CRs if the first CR duration was 1–2 years and 2/13 CRs if the CR was shorter). The system was similarly effective in stratifying the 137 patients who received both initial and salvage therapy at M.D. Anderson Cancer Center and the 69 patients who were referred in relapse (49% of whom had received >1 regimen for relapse prior to referral), with CR rates of 8/11, 11/24, 11/102, and 0/80, and of 3/4, 3/6, 12/58, and 0/18 being noted for the 4 groups in the M.D. Anderson Cancer Center and referred categories, respectively. Considering only patients who received conventional regimens (e.g., high-dose Ara-C, anthracycline, mitoxantrone, and etoposide), CR rates were 11/14, 14/22, 21/104, and 0/44, respectively, in the 4 groups. Applied prospectively to the 64 salvage AML, non-APL patients treated without allogeneic transplantation since 1995, the scheme's 4 groups had CR rates of 3/3, 4/14, 4/33, and 0/19, respectively. In patients treated from 1986 through 1989, CR rates were 8/11, 13/30, 23/135, and 0/74 in groups 1, 2, 3, and 4, respectively.

This four-group system could potentially facilitate comparison of new salvage regimens. Even with this system, however, there might be differences in CR rates in patients in the same group given the same treatment. In addition to the random variation introduced by small sample sizes, such differences could result from variations in the efficacy of post-first CR treatment, in the number of courses of salvage therapy given, and as a result of selection bias. Rarely do papers describing results of salvage regimens indicate that consecutive patients have been treated. For these reasons we believe that patients in phase II trials should be randomized to different treatments [7]. Although the number of patients enrolled will be inadequate to establish the superiority of one agent over another, as would be possible in a phase III trial, the randomization process will reduce bias. The four-group system described above could be used to stratify patients entering these randomized trials.

Selection of drugs for phase II trials

Once the maximum tolerated dose of a new drug has been determined in a phase I trial, it becomes important to establish whether the drug is sufficiently effective to warrant inclusion in a randomized comparative phase III trial. The phase II trial is the vehicle typically employed to address this issue. However, problems arise when several drugs are available for phase II testing, the preclinical rationale for each appears equally convincing, and the number of patients available is fewer than the 14–50 often required to complete the standard phase II trial. In this case the choice of which drug to test, i.e., the pre-phase II selection process, is typically carried out informally on the basis of in vitro data, toxicity study data, phase II results for similar or related treatments, and clinical judgment.

A design proposed by Thall and Estey [8] is essentially a statistical formalization of this informal process, with the important differences being that prior opinion regarding efficacy is expressed explicitly, patients are randomized among treatments, and the selection rule is based formally on clinical data. Prior opinion about the effectiveness of new drugs is encapsulated in a "prior probability distribution" (prior; a Bayesian concept) in which the physician assigns each CR probability (if this is the outcome of interest) a weight reflecting how likely he believes that probability is. The two important properties of a prior are the mean and the dispersion about the mean. The mean (P_0) and the dispersion can be thought of as the CR probability thought most likely by the clinician and the variance about the mean, respectively. For example, if the new drugs are to be investigated in a population where standard therapy produces a 20% CR rate, the mean would be 0.2. If the physician believes a priori that a CR rate well below or above 20% is likely, the dispersion will be broad. If he believes that the CR rate is likely to be concentrated around 20%, the dispersion will be narrow. In the Thall-Estey design the dispersion of the prior is very broad. This reflects the lack of experience with the new drugs. If the dispersion were narrow, the question of why the study is being performed might legitimately arise. The design also re-

Table 1 Design parameters and operating characteristics[a] (P_0 Standard [historical] treatment CR probability, k number of treatments evaluated, n maximal number of patients/treatment arm, PNS_0, probability that no treatment is selected [all $P = P_0$], PET_0 probability that a given arm is terminated early [$P+P_0$], EN_0 expected sample size when all $P = P_0$, PCS probability that the best treatment is selected [one $P = P_0+02$, all other $P = P_0$])

Stratum[b]	P_0	k	n	π_1	π_2	PNS_0	PET_0	EN_0	PCS
1	0.4	2	10	0.9	0.8	0.69	0.31	16.6	0.6
2	0.1–0.2	3	15	0.9	0.8	0.64	0.60	27	0.74

[a] Note that in frequentist statistical terminology, 1−PNS is the probability of a false-positive conclusion, whereas 1−PCS is the probability of a false-negative conclusion. However, these false-negative probabilities must be compared to the false-negative probability arising if one of k regimens is arbitrarily selected for future study. Assuming a priori that each treatment arm is equally likely to be successful, this is $1-1/k$. Stopping and acceptibility rules are as follows, based on x_m = number of CRs among the first m patients: if P_0 is 0.4 a treatment arm should be terminated if x_m/m is $\leq$0/3 or 1/7 and is acceptable for phase II study if x_m/m is $>$6/10; if P_0 is 0.1–0.2 a treatment arm should be terminated if x_m/m is $\leq$0/5 and is acceptable for phase II study if x_m/m is $>$3/15

[b] Strata are defined in the text and in Estey et al. [1]

quires that the priors for each new treatment be identical because the design randomizes patients among treatments relying on the a priori assumption that the treatments are equally effective.

Once the priors have been established, response data from each patient (in this example, CR or no CR) are incorporated into the prior to yield a posterior probability distribution (posterior), also characterized by a mean and dispersion. Thus, the posterior can be thought of as a function of both the prior and the observed data. The more disperse the prior, the more the posterior will reflect the observed data, and vice versa.

An important requirement is that any treatment exhibiting poor results be terminated early. Assume that the mean of the prior (P_0) is 0.2. Then, the posterior (i.e., current) probability that the CR rate with a given treatment arm is $<$20% can be determined by measuring the area under the posterior curve between 0 and 20%. Once this posterior probability (π_1) becomes very high, e.g., $>$90%, accrual in the arm stops.

If accrual in a given arm does not stop, a fixed number of patients (n) are entered on each arm. In the notation, k refers to the number of treatment arms being evaluated and, thus, $n\times k$ is the maximal number of patients that can be accrued. Assuming that n patients are accrued on >1 treatment arm, a treatment arm would be selected for future phase II testing provided that (1) it had a higher CR rate than the other arms and (2) it met a minimal acceptability criterion.

The minimal acceptability criterion is that the π_1 that the CR rate is $>P_0$ is high. This particular π_1 value is referred to as π_2. If each of two arms had the same CR rate and met the minimal acceptability criterion, both would be selected for a phase II study, which would be done by randomizing patients between the two treatments.

The exact values for π_1, π_2, n, and k are specified by the physician and statistician by examination of the operating characteristics that follow if particular values for these parameters are assumed. The operating characteristics of interest are the probability of early termination if a given arm is identical to the historical treatment (PET_0); the probability of not selecting any arm for future phase II study, given that all the treatment arms are identical to the historical treatment (PNS_0); and the probability of (correctly) selecting an arm for phase II study, given that it is superior by a fixed increment, e.g., 20%, to the historical treatment while all other treatments are identical to the historical treatment (PCS). PET_0, PNS_0, and PCS values for the proposed design are shown in Table 1, as are values for n, k, π_1, and π_2.

Given that a relatively small number of patients participate in the pre-phase II design described above, it is distinctly possible that a disproportionate number of favorable or unfavorable patients could be assigned to a particular treatment. To avoid the attendant possibilities of false-negative as well as false-positive results, the pre-phase II design is applied separately to each of two distinct prognostic groups (strata 1 and 2).

Two issues arise regarding design implementation. First, is it sensible to discontinue a regimen when, as the design calls for, none of the first three patients responds [the case in the 40% prognostic stratum in Table 1, 40% referring to the standard (historical) success probability] or when none of the first five patients responds (10–20% prognostic stratum)? Physicians are more familiar with designs such as that of Gehan [2], which call for stopping in a 10–20% group only if none of the first 14–29 patients responds, resulting in a false-negative rate of only 5%. Stopping if none of the first five patients responds increases the false-negative rate, e.g., to 26% in the scenario illustrated in Table 1 where two of three drugs have true response rates of 10% (the standard) but one drug has a true response rate of 30%. The justification for the increase obtained in the false-negative rate with the Thall-Estey design is essentially that (1) the worst false-negative scenario occurs when a drug is never tested and (2) the design enables more new drugs to be tested than does the standard design, at the price of a higher rate of false negativity. Whether physicians will, or should, use the Thall-Estey design might depend on their a priori, i.e., preclinical, assessment of the likelihood of success with various new agents. If one drug is believed to be much more promising than others, phase II designs with low false-negative rates should be chosen. If, on the other hand, distinctions between several new agents cannot be made a priori, then the Thall-Estey design is much more likely than standard designs to ensure that more drugs are tested, i.e., to avoid the scenario where a potentially useful drug never reaches clinical trial. Another factor involved in

the selection of a design is the number of available patients. If the number is large the use of a standard phase II design would permit a number of drugs to be tested.

References

1. Estey E, Kornblau S, Pierce S, Kantarjian H, Beran M, Keating M (1996) A stratification system for evaluating and selecting therapies in patients with relapsed or refractory AML. Blood 88:756
2. Thall PF, Estey E (1993) A Bayesian strategy for screening cancer treatments prior to phase II evaluation. Stat Med 12:1197
3. Hiddemann W, Martin WR, Sauerland CM, Heinecke A, Buchner T (1990) Definition of refractoriness against conventional chemotherapy in acute myeloid leukemia: a proposal based on the results of retreatment by thioguanine, cytosine arabinoside, and daunorubicin (TAD 9) in 150 patients with relapse after standardized first line therapy. Leukemia 4:184
4. Kornblau SM, Gandhi V, Andreeff M, Beran M, Kantarjian H, Koller CA, O'Brien S, Plunkett W, Estey E (1996) Clinical and laboratory studies of 2-chlorodeoxyadenosine ± cytosine arabinoside for relapsed or refractory acute myelogenous leukemia in adults. Leukemia 10:1563
5. List AF, Spier C, Greer J, Wolff S, Hutter J, Dorr R, Salmon S, Futscher V, Baier M, Dalton W (1993) Phase I/II trial of cyclosporine as a chemotherapy-resistance modifier in acute leukemia. J Clin Oncol 11:1652
6. Santana VM, Mirro J Jr, Kearns C, Schell MJ, Crom W, Blakely RL (1992) 2-Chlorodeoxyadenosine produces a high rate of complete hematologic remission in relapsed acute myeloid leukemia. J Clin Oncol 10:364
7. Simon R, Wittes RE, Ellenberg SS (1985) Randomized phase II clinical trials. Cancer Treat Rep 69:1375
8. Gehan E (1961) The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis 13:346