Shortcomings in the clinical evaluation of new drugs: acute myeloid leukemia as paradigm

Roland B. Walter,1,2 Frederick R. Appelbaum,1,3 Martin S. Tallman,4 Noel S. Weiss,5,6 Richard A. Larson,7 and Elihu H. Estey1,2

1Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 2Department of Medicine, Division of Hematology, University of Washington, Seattle, WA; 3Department of Medicine, Division of Medical Oncology, University of Washington, Seattle, WA; 4Memorial Sloan Kettering Cancer Center, Weill Cornell Medical College, New York, NY; 5Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; 6Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA; and 7Section of Hematology/Oncology, Department of Medicine, University of Chicago, IL

Drugs introduced over the past 25 years have benefitted many patients with acute myeloid leukemia (AML) and provided cure for some. Still, AML remains difficult to treat, and most patients will eventually die from their disease. Therefore, novel drugs and drug combinations are under intense investigation, and promising results eagerly awaited and embraced. However, drug development is lengthy and costs are staggering. While the phase 1–phase 2–phase 3 sequence of clinical drug testing has remained inviolate for decades, it appears intrinsically inefficient, and scientific flaws have been noted by many authors. Of major concern is the high frequency of false-positive results obtained in phase 2 studies. Here, we review features of phase 2 trials in AML that may contribute to this problem, particularly lack of control groups, patient heterogeneity, selection bias, and choice of end points. Recognizing these problems and challenges should provide us with opportunities to make drug development more efficient and less costly. We also suggest strategies for trial design improvement. Although our focus is on the treatment of AML, the principles that we highlight should be broadly applicable to the evaluation of new treatments for a variety of diseases. (Blood. 2010;116(14):2420-2428)

Introduction

Acute myeloid leukemia (AML) comprises a heterogeneous group of neoplasms characterized by an accumulation of clonal myeloid progenitor cells that do not differentiate normally.1,4 These disorders remain difficult to treat, and most patients will die of their disease within 1-2 years of diagnosis.1,2,5 Nonetheless, drugs introduced over the past 25 years, such as all-trans retinoic acid (ATRA), arsenic trioxide (ATO), and gemtuzumab ozogamicin (GO), have improved survival and the prospect of cure in some patient subsets.5-8 These successes have increased interest among patients and physicians, as well as investors and the general public, in identifying “promising” new drugs and drug combinations. Trials of such therapies are being conducted at ever accelerating rates.

The sequence of trials for clinical drug development has remained inviolate for decades.9 Phase 1 studies, typically conducted in patients with advanced or treatment-refractory disease, establish a maximum-tolerated or, more recently, an optimal biologic, dose for phase 2 trials. The latter primarily tests for a suggestion of efficacy. Rather than survival or quality of life, arguably the outcomes of most interest to patients, the criteria for efficacy frequently are presumed surrogates of benefit such as, in the case of AML, the achievement of complete remission (CR), CR with incomplete platelet recovery (CRP), or CR with incomplete blood count recovery (CRi).10,11 Furthermore, efficacy is often defined without reference to what might have been expected had similar patients received older, more standard therapies. Such a comparison is commonly delayed until phase 3 of clinical testing, in which large numbers of patients are randomized between newer investigational and more standard therapies. In contrast to earlier trials, phase 3 studies routinely address survival rather than response. These characteristics render phase 3 trials the key vehicle for regulatory drug approval.

Various authors have noted scientific flaws with the phase 1–phase 2–phase 3 paradigm.12-16 Here, we are principally concerned with the inefficiency engendered by standard means of conducting and reporting phase 2 studies. The sequential trial scheme puts major emphasis on such studies because they typically inform the decision to proceed to a phase 3 evaluation. Unfortunately, current approaches to the conduct of phase 2 studies have substantial shortcomings. Particularly as more drugs are introduced, these flaws delay evaluation of new drugs and increase their cost. While we address AML to elucidate some problems with phase 2 studies and to suggest some solutions, we believe our points are generally applicable.

The cost and inefficiency of oncology drug development

The cost of the current drug development process is well documented. In 2003, a landmark study randomly selected 68 drugs from a proprietary investigational compound database and estimated an average of $802 million (in 2000 US dollars) in research and development costs to bring a new chemical entity to market.17 A very similar average estimate, $868 million, was obtained using a public database of new drugs entering initial human clinical trials...
between 1989 and 2002. Estimates ranged from $500 million to more than $2 billion, with drugs against blood disorders or cancer being more costly, averaging $906 million and $1.04 billion, respectively.\textsuperscript{18} The methods to calculate these estimates and the costs themselves have been questioned, with some contending that drug companies might overstate costs.\textsuperscript{19,20} Regardless, there is little doubt or controversy that drug development is staggeringly expensive.

The US Food and Drug Administration recently estimated that only approximately 8% of new medicinal compounds entering phase 1 testing will reach the market, reflecting a worsening outlook from the historical figure of approximately 14%.\textsuperscript{21} The likelihood of approval is probably even lower for oncology drugs (around 5%–10% in estimates from 1975 to 2000).\textsuperscript{22,23} A striking feature of the current clinical drug testing process is the difficulty, at any point, of predicting ultimate success of a novel candidate drug.\textsuperscript{21} Relative to other drugs, anticancer agents are more likely to progress to phase 3 testing.\textsuperscript{24} However, even drugs that enter phase 3 are frequently found to be no better than the standard control therapy. This is particularly true for cancer drugs, which have a notably lower likelihood of success in phase 3 than other drugs, with recent estimates as low as 41%.\textsuperscript{23,24} Phase 3 trials that do not obtain positive results are problematic: commonly enrolling hundreds to thousands of participants, these trials are generally much more laborious, time-consuming, and expensive than phase 1 or phase 2 studies.\textsuperscript{17,18} Recent studies estimate an average cost of $1 million for publicly funded, and $10 million for pharmaceutical industry-funded, phase 3 trials, with an average of 4.5 years to completion.\textsuperscript{25,26} The high probability of negative results in phase 3 studies leads to inefficiencies and delays in testing new drugs which, due to financial and logistical constraints, can only enter phase 3 after previous drugs have completed the phase 3 process. Since the phase 2 trial acts as gatekeeper for the phase 3 trial, we and others\textsuperscript{24,25} believe that more attention should be given to the conduct of phase 2 trials to minimize the risk of overly optimistic results and to limit the number of subsequent negative phase 3 trials.

Limitations of phase 2 studies in identifying useful new therapies

Reports of early phase trials of new therapies are presumably read in anticipation that positive results herald therapeutic advances. Such reports include abstracts presented at major meetings and more mature, possibly more influential, final manuscripts published in peer-reviewed scientific journals. Regardless of the publication category, the predictive value of a positive report for subsequent clinical utility is low, at least for anti-AML drugs. In 2006, we addressed this issue using abstracts submitted to the Annual Meeting of the American Society of Hematology (ASH) between 1993 and 2001.\textsuperscript{27} Specifically, we reviewed all abstracts if they reported on an early phase trial of a new drug used alone or in combination for adults with AML other than acute promyelocytic leukemia (APL). The year 2001 was chosen to allow a minimal follow-up of 5 years. PubMed, a database of the US National Library of Medicine, was then used to identify subsequent AML-related studies using these drugs and/or drug combinations. Sixty-three of the 91 abstracts (69%) involving 37 separate drugs were judged positive, based on conclusions that the therapy was “active,” “promising,” “worthy of further investigation,” and so on; only 14 abstracts (15%) were considered “negative,” while 14 were regarded as “inconclusive.” Forty-five of 63 (71%) positive abstracts covering 27 of the 37 separate drugs subsequently appeared in peer-reviewed journals: the positive conclusion was unaltered in 38 of the 45, whereas it was changed to negative in 5 and to inconclusive in 2. Only 3 of the 37 positive drugs were later found to be positive in a randomized phase 3 trial (GO, interleukin-2/histamine, cyclosporine/infusional daunorubicin/cytarabine), and only GO has migrated into clinical practice, although usually not for the indication suggested by the randomized trial. 4 of the drugs found positive in early studies yielded negative results in randomized trials. Importantly, the majority of drugs (30/37, 81%) with “positive” early phase trial data remained unevaluated in randomized studies, possibly because the quality of data did not appear trustworthy enough to justify investment in phase 3 or because of limited resources, and are not used in clinical practice.\textsuperscript{27} More recently, we found that positive early phase AML drug evaluations, published in peer-reviewed journals between 1989 and 2003, also most often do not lead to subsequent randomized studies (R.B.W., unpublished observation, May 2010).

The observation that promising results from phase 2 studies do not translate into positive phase 3 trials is not restricted to AML. For example, Zia et al reviewed phase 3 trials in advanced solid malignancies published between 1998 and 2003 and identified 43 that used identical therapeutic regimens as 49 previously conducted phase 2 studies.\textsuperscript{28} Only 12 of the phase 3 studies were considered positive, and in 81% of the phase 3 trials, the degree of clinical response was lower than in the preceding phase 2 trials. Somewhat more optimistic were the findings from a review of all phase 3 studies of biologic agents against advanced cancers published from 1985 to 2005: among 351 phase 2 studies, 167 subsequent phase 3 trials were positive.\textsuperscript{29} Still, the majority of phase 3 trials are negative.

Predictors of a successful phase 3 study

Given the costs of phase 3 trials and the observation that many phase 2 trials never advance to phase 3 (or are negative when they do advance), it is imperative to identify the phase 2 trial characteristics that predict a positive phase 3 study. Attention to such characteristics in trial design may help optimize drug development and minimize the resources expended on drugs that will likely fail in later stages of drug testing. Systematic analyses of factors predictive for successful translation to phase 3 trials have not been conducted in AML or other hematologic malignancies, and only limited information is available for other cancers. In their review of phase 3 studies of conventional chemotherapeutics in advanced solid malignancies, Zia et al identified sample size of the phase 2 trial as the only variable possibly associated with a positive phase 3 study, while multicenter trial conduct, randomization, frequency of response, and journal impact factor were not.\textsuperscript{28} In contrast, in phase 2 studies of targeted agents such as antibodies, immunotherapeutics, oncolytic virotherapies, biologic response modifiers, and small molecule inhibitors, factors predictive of success in subsequent phase 3 trials included multicenter trial conduct, positivity of the phase 2 trial (a minority of phase 3 trials followed negative phase 2 studies), and a trial conducted by a pharmaceutical company (89.5%, vs 44.2% for academic, 45.2% for cooperative group, and 46.3% for research institute trial).\textsuperscript{29}
**Problems with, and some suggestions for, phase 2 trials in AML**

Absent formal studies, elucidation of features of phase 2 trials in AML that make it unlikely that a positive result will be reproduced in phase 3 (or that a phase 3 trial will even be conducted) must remain speculative. However, the following common features of AML phase 2 trials are likely relevant (Table 1).

**Small study size**

The issue of small study size is best illustrated by Leopold and Willemze’s review of trials in refractory and relapsed AML, a frequent setting for testing of new agents. The authors identified 112 peer-reviewed reports describing treatment of AML in first relapse published between 1979 and 1999. Only 31 (28%) of these enrolled at least 20 primarily adult patients and provided information on duration of first CR, a critical predictor of response to such salvage therapy. Seventeen of the 31 reports were prospective single-arm phase 2 trials with a median size of 26 patients. This number is considerably less than that specified in standard phase 2 designs and consequently increases the probability of false-positive or false-negative results unless the new drug is greatly superior or inferior to historical treatment. While false-positive results will be corrected in subsequent phase 3 trials, false-negative results are at least equally as problematic because there would be little incentive to re-test the drug. Consequently, a valuable therapy might be lost forever.

Less obvious is the issue of small historical sample size. One problem with some historically controlled studies stems from their (incorrect) assumption that response data are derived from an infinitely large control group. Because this is usually not the case, the historic response itself represents an estimate characterized by a mean and variance. Under these circumstances, the actual probability of obtaining a false-positive result (type I error) is higher than the nominal rate, with the error magnitude inversely related to the size of the historical control group. A 1-arm design with historic control may thus only be preferred for a phase 2 study when limited patients are available and the historical response proportion is well-established, whereas a 2-arm design with randomization may be preferred with larger sample sizes or if the uncertainty in the historical degree of response is large. To properly interpret the results of phase 2 studies, it follows from the above that scientific reports should, but often fail to, specify the false-positive and false-negative errors associated with number of patients treated; furthermore, for studies that use the experience of previously untreated or differently treated patients as a basis for comparison, the number of such patients along with their distribution of clinical characteristics should be provided.

**Lack of a control group: the potential value of randomized phase 2 trials**

A major shortcoming of phase 2 studies in AML is that, often, no reference to a control group is made, rendering it difficult to estimate how good results truly are. Thus, while such trials may suggest that the new agent or combination is active, it remains unclear whether it is better than an older therapy, which is also active. Fundamentally, phase 2 studies are inherently comparative. New treatments are seldom good or bad, but rather better or worse than some standard, and medical decision-making essentially involves choosing the treatment with the highest benefit-to-risk ratio. Nevertheless, the review by Leopold and Willemze and considerable experience suggest that the great majority of phase 2 studies in the AML literature lack even an historical control group.

We believe that the chance of an apparently promising phase 2 study not being confirmed in phase 3 is reduced by inclusion of a control group, be it historical or concurrent (nonrandomized or randomized). Although the former requires fewer new patients, its historical nature makes it difficult to accurately specify the rate of no interest (null hypothesis) that will be compared with the rate seen with the new agent. In the absence of any true benefit with the new agent, underestimation of the null can drastically increase the probability of proceeding to a phase 3 trial. As a result, the number of patients treated with the new agent will likely be much larger than suggested by the nominal phase 2 sample size, although the new agent offers no benefit over the historical treatment. Conversely, overestimation of the null can dramatically reduce the probability of continuing to a phase 3 trial, even when the new treatment provides benefit over the historical treatment. The imprecision in historical estimates can thus have an important effect on the study, and statistical methods have been developed that aim to adjust for this uncertainty.

Problems in comparing historical responses (or, for that matter, responses of contemporary nonrandomized controls) with those seen in a phase 2 evaluation of a new drug arise from 2 fundamental sources. First, the 2 groups may differ in the distribution of known
shortcomings in aml drug evaluation

predictive factors. These include, for example, duration of first CR in relapsed/refractory AML and cytogenetics in untreated older patients. Multivariate analyses can be used to adjust the comparisons between new and control data to account for differences in important known prognostic factors and reduce the risk of relevant confounding. However, this approach ignores the second difficulty in comparing historical and current responses, namely the presence of unrecognized and hence unmeasured latent variables. They can serve as confounding variables and lead to erroneous estimates of the impact of the new therapy. Tang et al categorized such variables into those resulting from patient temporal drift or patient selection effects. The former is a systematic, population-wide shift in outcomes, for example, caused by changes in disease diagnosis or classification, or efficacy of supportive and later-line therapies, whereas the latter is a difference in the patient population of a trial compared with the population enrolled in historical control trials, for example, with respect to patient or institutional characteristics. The probability of a false-positive finding can increase dramatically even with modest patient temporal drift or patient selection effects; this error is not reduced, and in fact may be even made worse, by increases in the sample size in the contemporary phase 2 trial. Although methods to account for imprecision in historical estimates have been proposed, current phase 2 cancer trials typically do not contain such statistical adjustments. Moreover, many phase 2 studies fail to cite the source of historical data used; these trials are more likely to declare an agent to be active.

Although the effect of confounding variables can be estimated, the best way to account for them is through randomization, if possible. While not all scientific questions need to be addressed with a randomized control trial (RCT), there is little disagreement that RCTs constitute the most rigorous and best method of evaluating the efficacy of a therapeutic intervention. By comparison, the opinions regarding validity of information inferred from nonrandomized, observational studies vary widely. Due to the inherent propensity to introduce biases, some experts see considerable dangers to clinical research if observational studies replace RCTs. Indeed, a systemic comparison of RCTs and historic control trials for therapies that were studied by both methods indicated that historic control patients generally do worse than the control group from the RCT. This would suggest that historic control trials are systematically biased toward favoring the new therapy. Other reports came to similar conclusions. On the other hand, some authors concluded that observational studies may provide valid information and may not consistently overestimate the magnitude of treatment effects.

Weiss proposed several criteria to assess the validity of comparisons from nonrandomized trials. These include the requirements that illness is monitored similarly among the treatment groups, and that baseline differences in prognostic factors are small (or can be made small by statistical adjustments) relative to the size of the observed difference in outcome. Possibly because of the difficulty in ascertaining whether these criteria have been met, randomized phase 2 trials, that is, randomized studies with far fewer patients than traditional phase 3 trials, have been advocated and are increasingly frequent in oncology. Indeed, it has been estimated that 24% to 30% of phase 2 studies in oncology are randomized. A recommendation for additional study was made in 45% of the 266 randomized phase 2 oncology trials reviewed by Lee and Feng. In contrast, such a recommendation was made in 75% of all phase 2 oncology studies. Four of the 6 phase 3 studies that followed a randomized phase 2 trial were reported as positive. Although the sample size is small, this higher-than-expected proportion suggests that randomization in phase 2 could decrease the number of negative phase 3 trials.

For comparable theoretical statistical operating characteristics, randomized designs generally require up to 4 times as many patients as single-arm studies with historical controls. Although this increase in sample size may be justifiable if it reduces the likelihood of subsequent negative large phase 3 studies, it may offer a challenge to timely study completion in a disease with limited patients. This led to attempts to develop randomized trial designs that require fewer patients but still protect against some of the potential shortcomings of single arm studies. Several designs for randomized phase 2 trials have been proposed, including selection (pick-the-winner) trials, screening trials, and randomized discontinuation trials. Medical Research Council (MRC) trials for older patients with untreated AML now employ such a pick-the-winner design. In these trials, the goal is to select the therapy with superior response for further testing. To address the concern that one therapy would be chosen even if none was superior to an established standard therapy, each arm of the selection design can be constructed as a 2-stage design to be compared separately against a historical control. For selection trials, controlling false-positive errors is less relevant than controlling false-negative errors as this trial aims to ensure that there is a high probability that a regimen is selected if it indeed is superior. In other words, the selection design addresses the view that the worst false-negative occurs if a new treatment is not studied at all. Nevertheless, because of their size, randomized selection trials are underpowered for performing formal hypothesis testing or comparisons of primary or secondary end points across selection arms, and have therefore been the subject of criticism, in particular for being prone to false-negative conclusions. However, while selection trials commonly have a statistical power that is less than the 80% conventionally used in a phase 3 trial, the figure of 80% ignores the possibility that the new agent was selected informally among, for example, 4 possible new agents. History suggests that preclinical rationale is often insufficient to know a priori which of the 4 new agents is best. The selection design can thus be useful to pick the likely most effective drug for further study; however, this design does have an undesirable high false-positive rate consequent to its small size. Therefore, while helpful in circumstances where there is uncertainty as to the relative value of a multitude of new treatments, a selection trial must be followed by a confirmatory trial.

Information from a randomized phase 2 trial can be used in an integrated or seamless phase 2/3 trial; these designs allow phase 2 patient data to be used in the principal phase 3 trial analysis and thus reduce the number of patients needed for phase 3. Such trials can be very flexible and monitor either response or survival as primary end point in phase 2, and they can test multiple experimental arms and/or a concurrent randomized control. While there are some limitations with these designs, such as the requirement for relatively large sample sizes in the phase 2 portion, they substantially reduce the sometimes very long delays between completion of the phase 2 and initiation of the phase 3 study. In this connection it is noteworthy that a median of 784 and 808 days passed from initial conception of the study to activation of recent phase 3 cooperative group trials of the Cancer and Leukemia Group B (CALGB) and Eastern Oncology Group (ECOG), respectively.
Patient heterogeneity: the problems of confounding and effect modification

Outcome variability is characteristic of AML. For example, among patients who have relapsed after a first CR, the likelihood of response—less than 5% to 60%—to second-line (salvage) therapy depends on the duration of the first CR, cytogenetics at diagnosis, age, and number of prior induction therapies.\(^{31,67}\) Numerous factors, principally cytogenetics, are predictive of prognosis among untreated older adults with AML, another group commonly enrolled in phase 2 studies. Despite this, the literature typically regards relapsed patients as a homogeneous group and does likewise for untreated older patients. This is problematic, as the interpretation of a new drug’s activity can be confounded by the particular composition of better and worse prognosis patients receiving the drug. In fact, the lack of a randomized study combined with the inclusion of an heterogeneous patient population regarding important prognostic factors, resulting in difficulties in the interpretation of study results, was a main reason that the FDA Oncology Drug Advisory Committee voted in 2009 against approval of clofarabine for the treatment of patients aged 60 and above.\(^{68}\)

Given the above, phase 2 studies should account for patient heterogeneity. The simplest method is to conduct distinct trials in various prognostic groups, for example patients with better and worse cytogenetics. Although preferable to averaging and considering such patients as 1 group, separate trials increase sample size and study duration. Thus, several methods have been developed for handling response heterogeneity within a phase 2 trial.\(^{69}\) Still, some of these methods, similar to the conduct of separate trials, do not formally allow results from 1 subgroup to influence trial conduct (eg, stopping or continuing) in another group. This is particularly problematic when treatment-subgroup interactions exist, that is, when a treatment has different effects in different prognostic groups. Wathan et al have proposed a hierarchical Bayesian design to address this problem.\(^{70}\) As in the case of separate trials, stopping rules are subgroup specific. However, in contrast to separate trials, the design examines accumulating data to see whether a given treatment might have similar effects in different prognostic groups and allows data from 2 groups to be combined to the extent that such borrowing of strength is justified by these data. Although the design is computationally complex, advances in computing algorithms and in computing power will likely facilitate use of these and other trial designs. Formal methods for subset analysis might prevent early access of children to beneficial therapies.

Another group of patients that is often excluded from AML trials are children and adolescents, although the need for clinical trials in these patients is increasingly recognized by the scientific community and oftentimes required for drug approval by regulatory authorities. While the policy of testing new drugs only after they are evaluated, and sometimes approved, in adults protects children from ineffective drugs and unwanted drug toxicities, this may prevent early access of children to beneficial therapies.

Choice of study end point(s)

Currently, the most commonly used primary end point in phase 2 trials is probability of response (response rate). Reasons of economy and time lead to the selection of a primary end point that is not one of greatest interest to patients and regulatory agencies, for example, survival or improvement in quality of life, but rather a more common antecedent (surrogate) such as response.\(^{50,73}\) In AML, the choice of response rate as end point is problematic as most responses are transient and may add little prolongation of survival time. An emphasis on response duration, relapse-free survival (RFS; also called disease-free survival [DFS]), or overall survival (OS) may obviate this problem. Unlike RFS or OS, response duration is subject to the competing risk of death without relapse. Because “relapse” and “death while still in response” are not mutually independent, the probability of remaining relapse free is thus not accurately estimated with the Kaplan-Meier method, and cumulative incidences of relapse should be calculated instead.\(^{30}\) Compared with OS, RFS is not confounded by receipt of subsequent salvage therapy; it also occurs earlier than OS, thus shortening the duration of the study and follow-up.\(^{57,58,74}\) Only recently, expert panels including those on behalf of the European LeukemiaNet\(^{11}\) or an International Working Group\(^{19}\) have provided...
recommendations for standard names and definitions for these and
other outcome measures. Standardized response criteria and sur-
vival outcomes that are widely accepted and employed will
undoubtedly facilitate the interpretation of clinical studies.

Response duration and response are both used as surrogates for
survival. Studies using such surrogate end points can be convincing
if there is reason to believe that the surrogate lies on a pathway
linking the treatment and the more relevant outcomes of survival or
quality-of-life, as is the case for CR or CR duration in AML. Use
of surrogate end points may increase efficiency to the extent that
they occur relatively more commonly or more quickly. In fact,
conditional or accelerated drug approval in the United States can
use phase 2 data and rely on a surrogate that is likely to benefit
patients directly while further studies demonstrating direct patient
benefit are under way. Yet, caution is necessary in the choice of
surrogates and interpretation of results from trials with surrogate
end points, and there are many prominent examples where such
trials have been misleading. To be completely valid for the
assessment of effectiveness, a surrogate end point must fully
capture the effects of treatment on the clinical end point; this
requirement is very difficult to satisfy in practice.

The superiority of response duration over response as a
surrogate illustrates that not all surrogates are equal. Another
example contrasts CR with lesser degrees of response. Almost
50 years ago, Freireich et al demonstrated that patients with AML
who achieve a CR live longer than those who do not, with the
difference in survival accounted for by the time spent in CR. Recent
years have seen a broadening of response categories.
Specifically, in 2003, new categories of responses were proposed,
including CRi and the closely related CRp. We have very recently
demonstrated that, after adjustment for covariates, the RFS of
patients achieving CR was longer than that of patients achieving
CRp, whereas patients with CRp survived longer than those with
resistant disease. These data indicated that CR is of particular
clinical significance and should be reported as separate response in
AML. Nonetheless, these findings also validated CRp as a clini-
cally meaningful response. On the other hand, the effect on survival
of CRi, presumably a lesser response than CRp, or of various
categories of hematologic improvement (HI) is unknown. While
inclusion of CRi or HI will increase the overall probability of
response to new drugs in phase 2, subsequent phase 3 studies may
find that the new drug does not improve survival simply because
some of the tabulated responses have no influence on survival.
Certainly, efforts to examine the relation between responses such as
CRi or HI and survival should be encouraged.

The choice of the most appropriate end point is critical for the
study design. Although response rates in most phase 3 trials are
lower than those in preceding phase 2 studies and are not predictive
of a positive phase 3 trial, response-based end points are still
relevant and may be appropriate, in particular for early phase 2
studies. However, for later phase 2 studies, and possibly for earlier
ones, end points such as remission duration, RFS, and OS should be
considered as primary. Furthermore, the use of such end points
earlier in clinical drug testing may conceivably reduce the number
of negative phase 3 trials. Survival as primary study end point
becomes particularly relevant in view of increasing examples of
anticancer therapeutics resulting in prolongation of RFS or OS with
very modest tumor responses or in patient categorized as nonre-
responders. Azacitidine may be 1 example of such a therapeutic in
elderly patients with AML. Use of response duration or survival measures as study end
point mandates attention to timing of follow-up tests, particularly
bone marrow examinations. A standardized approach to monitoring
will improve inter-study comparisons and reduce the variability in
outcome assessments that are introduced because of differing
monitoring schemes. In APL, sequential postremission disease
assessment is recommended for some patients and offers the
opportunity for preemptive therapy to prevent disease progression
and overt morphologic relapse if minimal residual disease (MRD) is
detected. In contrast, sequential postremission bone marrow
examinations are often not performed at standardized intervals in
non-APL AML; in fact, almost 15 years ago, Estey and Pierce
concluded that there was no clinical benefit to routine bone marrow
examination (typically every 2-4 months) in patients in remission
rather than obtaining bone marrow studies only should blood
counts worsen. Although often accepted as clinically reasonable,
this policy likely overestimates remission duration and RFS and
appears appropriate for clinical trials only if peripheral blood
counts are obtained at standard times and if uniform criteria are
used to define blood count deterioration. Nevertheless, it seems
preferable to adopt a standardized approach for bone marrow
monitoring, for example, every 3 months for 2 years and every
6 months for the following 2 to 3 years. This is particularly true as
monitoring moves beyond morphology to encompass detection of
minimal residual disease (MRD) by flow cytometric or molecular
assays, which are increasingly recognized as sensitive and specific
indicators of eventual morphologic relapse. With emerging evi-
dence that MRD monitoring helps optimize postremission therapy
and may lead to improved outcome of patients with non-APL
AML, it is likely that serial disease monitoring will become
more broadly accepted. However, the ultimate clinical utility of
MRD monitoring will depend on the development of better
therapies for relapse or conclusive demonstration that treatment at
time of detection of MRD rather than at morphologic relapse
improves clinical outcome; the latter is certainly a testable
hypothesis.

Of particular consideration in the design of AML trials is the
impact of allogeneic hematopoietic cell transplantation (HCT) as a
salvage or consolidation therapy. In many instances, experimen-
tal salvage (or induction) therapies are administered with the intent
of cytoreduction and transplantation as quickly as possible. Al-
though this strategy may be of benefit for individual patients, it
interferes with the ability to assess achievement of CR as a
surrogate measure of response as well as duration of CR and
survival after administration of the new agent.

Early publication

Rowe et al used an analysis of 6 consecutive clinical trials from the
ECOG to highlight problems with premature reporting of data. Reporting survival data 3 years after completion of study accrual
appeared to reflect mature study data accurately. In contrast,
while treatment results presented 1 year from conclusion of study
accrual were unlikely to be completely contradicted by further
follow-up, some differences in survival measures were noted,
warranting caution in the interpretation of data and their com-
parison with other published reports. More problematic are
studies presented upon or even before completion of study
accrual. These findings have clear implications for journal editors
and reviewers.
Subgroup-focused trials based on presumed mechanism of action of study drug

Important opportunities for drug development are likely to occur consequent to the increasingly frequent identification of cytogenetic and molecular markers in AML. They have refined our ability to provide prognostic information and risk determination for subgroups of patients and have helped in the development of subset-based treatment algorithms. Although ATRA and ATO in APL are the paradigm of such specific therapeutic guidance, it seems inevitable that similar instances will be found in non-APL AML. Current examples include high-dose cytarabine and GO in core-binding factor (CBF) AML, small-molecule inhibitors in AML with an internal tandem duplication (ITD) in the FMS-like tyrosine kinase-3 (FLT3) gene, and possibly ATRA in nucleophosmin (NPM1)–mutated AML and decitabine in AML with higher levels of miR-29h. Such markers may also serve to inform clinicians when to be more enthusiastic about the use of experimental rather than standard therapy, for example, in AML with monosomy karyotype, although such general guidance is less useful than the more specific guidance noted above. Of course, the success of future personalized approaches will depend on the demonstration of a correlation between ability to target a somatic abnormality and clinical response. Further refinement in the identification of patients likely to benefit from specific therapies may result from advances in pharmacogenomics. Subset-based approaches may increase the efficiency of the drug development process through shorter development times and smaller and/or fewer clinical trials, as fewer of these targeted patients will need to be enrolled in clinical trials to demonstrate clinical efficacy. This notion is well supported by experience with the humanized anti–HER-2 monoclonal antibody, trastuzumab, where restriction to patients with HER-2/neu overexpressing breast cancer allowed reduction of target enrollment from almost 2200 to 470 patients, reduced the duration of the clinical trial from estimated 10 years to 1.6 years, and saved an estimated $35 million in clinical trial costs. Nonetheless, subgroup-specific drug development is not without pitfalls, as demonstrated by the example of tipifarnib. Although believed to be specific for RAS mutations, no such specificity was observed in patients. Hence, the genetic and molecular diversity in AML may pose a formidable challenge to identifying adequately sized subgroups of suitable patients in whom to test specific therapies. If this challenge can be overcome, the large phase 3 trial with its assumption of an unrealistic amount of patient homogeneity may seem increasingly anachronistic.

Conclusions

Historically, some new therapies have been found to be effective in AML based on phase 2 results in a small number of patients and in the absence of even a historical control group. Nonetheless, we have illustrated that such cases are the exception, not the rule. Accordingly, the continuation of such practices in phase 2 is undesirable and limits the role of phase 2 studies as a gatekeeper to phase 3 trials. We believe that improvements would include larger phase 2 studies, inclusion of (preferably randomized) controls, consideration of integrated phase 2/3 studies, accounting for patient heterogeneity even in small randomized studies, provision of information about the number of patients available for study versus those actually treated, and avoidance of unvalidated surrogate end points and premature publication (Table 1). We are confident that attention to these matters will increase the efficiency and reduce the cost of new drug development both in AML and other diseases.

Acknowledgment

This work was supported by a grant from the National Cancer Institute/National Institutes of Health (P30-CA15704-35S6).

Authorship

Contribution: R.B.W. and E.H.E. were responsible for the conception and writing of the paper; and F.R.A., M.S.T., N.S.W., and R.A.L. contributed to the writing of the paper. Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Roland B. Walter, MD, PhD, Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, D2-190, Seattle, WA 98109-1024; e-mail: rwalter@fhcrc.org.

References

18. Adams CP, Brantner VV. Estimating the cost of


64. Koepsell TD, Weiss NS. Randomized trials. In:


Shortcomings in the clinical evaluation of new drugs: acute myeloid leukemia as paradigm