Review: Clinical Trial Methodology

CLINICAL INVESTIGATION

Considerations in the design of clinical trials for pediatric acute lymphoblastic leukemia

Clin. Invest. (2013) 3(9), 849–858

Acute lymphoblastic leukemia (ALL) is the most common pediatric malignancy. Although outcomes for children with ALL have improved dramatically over the last 50 years, ALL remains the leading cause of childhood cancer death. In addition, high-risk patient subsets can be identified with significantly inferior survival. In the current era of therapies directed at specific molecular targets, the use of conventional randomized Phase III trials to show benefit from a new treatment regimen may not be feasible when these biologically defined subsets are small. This review presents the traditional approaches to designing trials for children with ALL, as well as innovative approaches attempting to study the benefit of new treatments as reliably as possible for patient subsets with distinctive biological characteristics.

Keywords: acute lymphoblastic leukemia • clinical trials • historical controls

Meenakshi Devidas*[1] & James R Anderson[2]

[1]Department of Biostatistics, Colleges of Medicine, Public Health & Health Professions, University of Florida, 6011 NW 1st Place, Gainesville, FL 32607, USA
[2]College of Public Health, University of Nebraska Medical Center, 984355 Nebraska Medical Center, Omaha, NE 68198-4355, USA
*Author for correspondence:
Tel.: +1 352 273 0551
Fax: +1 352 392 8162
E-mail: mdevidas@cog.ufl.edu

Background

Acute lymphoblastic leukemia (ALL) is the most common pediatric malignancy, accounting for 25% of cancers occurring in children <15 years of age and 19% among those <20 years [1]. Outcomes for children with ALL have improved dramatically over the last 50 years, with 5-year overall survival (OS) now over 90% [2–4]. Figure 1 illustrates the improvements in survival from one era of trials to the next for patients on Children's Oncology Group (COG) trials for newly diagnosed ALL. However, despite steady improvements in outcome, because ALL is the most common childhood cancer, it remains the leading cause of childhood cancer death. In addition, high-risk subsets, including infants, newly diagnosed older children with high white blood cell count at diagnosis, patients with poor early response or with T-cell ALL, and all patients with relapsed ALL, have historically had significantly inferior outcomes. The COG is a National Cancer Institute sponsored clinical trials group with over 200 participating centers. COG captures a majority of children <15 years of age with cancer in North America on its clinical trials. Recent ALL clinical trials run by the COG have demonstrated improvements in outcomes for these patients by optimizing treatment with standard chemotherapy agents. It has also begun to study therapies developed on the basis of the presence of certain biomarkers and directed at specific molecular targets.

For the majority of children, the high cure rates have been achieved through risk-stratified therapy combining multiple chemotherapeutic agents. ALL is a heterogeneous disease comprised of morphologically identical leukemias arising from different biological mechanisms [5,6]. The COG has developed a risk stratification system for patients with newly diagnosed ALL that incorporates key clinical features including age at diagnosis, white blood cell count, immunophenotype, presence/absence of CNS or testicular leukemia, presence/absence of specific sentinel genetic

FUTURE SCIENCE part of fsg

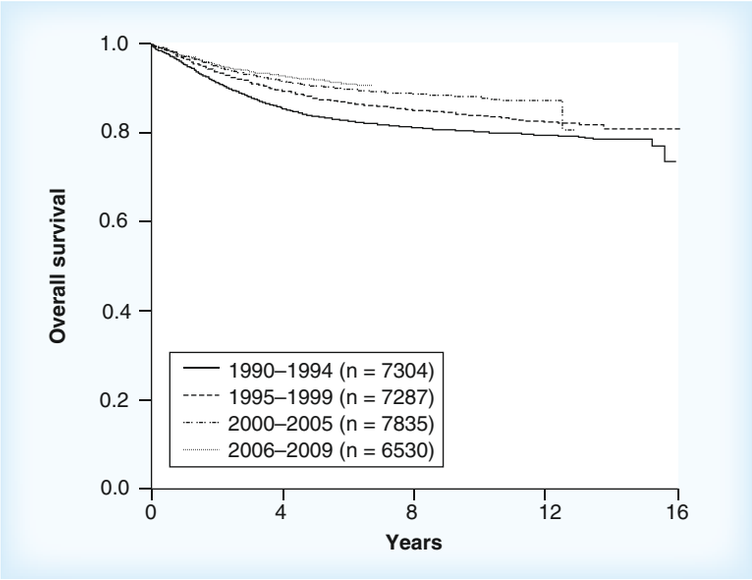

Figure 1. Overall survival by treatment era. Overall survival probability for acute lymphoblastic leukemia patients enrolled in Children's Oncology Group trials between 1990–1994, 1995–1999, 2000–2005 and 2006–2009. 5-year survival increased over time: 1990–1994: 83.7 ± 0.4%; 1995–1999: 87.7 ± 0.4%; 2000–2005: 90.4 ± 0.4%; and 2005–2009: 91.9 ± 0.6% ($p < 0.0001$).

lesions (good risk: *ETV6–RUNX1* fusion or hyperdiploidy with favorable chromosome trisomies; poor risk: *MLL*-rearrangements [*MLL*-R], hypodiploidy, intrachromosomal amplification of chromosome 21, Philadelphia chromosome positive [Ph$^+$] ALL), and early minimal residual disease (MRD) response for risk stratification and treatment allocation [7,8]. Other ALL clinical trials groups in Europe and elsewhere have developed similar risk classification systems directing therapy for these patients.

The process of risk stratification for treatment allocation is likely to lead to better outcomes for specific patient subsets, as shown by improved event-free survival (EFS) with the addition of imatinib into a treatment regimen for children with Ph$^+$ ALL [9]. Risk stratification for pediatric ALL in COG is complex, and is based on early clinical and biologic data in newly diagnosed patients. The presence of BCR–ABL1 fusion is considered an adverse prognostic factor; patients identified to have this are removed at an early stage from frontline trials for newly diagnosed ALL and are enrolled on the trial for Ph$^+$ ALL, which involves the use of a tyrosine kinase inhibitor in therapy. The recent discovery of JAK family mutations in a previously unrecognized subtype of high-risk B-precursor ALL has resulted in plans to include a JAK inhibitor into the treatment regimen for patients with JAK mutant ALL [10]. While the biological rationale is compelling for these new treatment approaches, the study design issues are significant. Cancer is uncommon in children compared with adults, and even in common pediatric cancers like ALL, studies are now being directed at patient subsets that may not have more than 5–10% of the whole ALL population. The conventional approach of using a randomized Phase III trial to show benefit from a new treatment regimen may not be feasible for some of these small, biologically defined populations. This review presents the traditional approaches to designing trials for children with ALL, as well as innovative approaches attempting to study the benefit of new treatments as reliably as possible for patient subsets with distinctive biological characteristics.

Phase I trials

Single agent Phase I studies for children are always performed after similar studies have been completed in adults and are generally performed in patients with solid tumors, and rarely in children with leukemia. The primary purpose of Phase I trials is to identify an appropriate dose of the new agent for further study in children, to determine if the pharmacokinetic parameters are similar to those observed for adults and to ensure that there are no serious toxicities that would prevent incorporation of the new agent in pediatric clinical trials. In general, children tend to tolerate anticancer agents at least as well as adults and so the recommended Phase II dose in children is often very close to that for adults. Since treatment regimens for pediatric ALL have similar criteria for hematologic dose-limiting toxicities and recovery as do regimens developed for solid tumors, the recommended Phase II dose in solid tumors is usually applicable to the ALL setting. Hence, leukemia-specific Phase I trials are unnecessary for agents being developed for ALL, in situations where a Phase I study of the agent will be conducted in children with solid tumors. For agents requiring dose escalation for pediatric ALL, generally the starting dose chosen is 70–80% of the adult maximum-tolerated dose (MTD) with an escalation to a maximum of one or two dose levels above the adult MTD. The '3 + 3' design is the standard used in dose-escalation studies, where the MTD is defined as the highest dose at which no more than one out of six patients experience a dose-limiting toxicity as the MTD [11]. Other Phase I trial designs include the 'rolling six design' and the modified continual reassessment method [12,13]. A dose-escalation study may not be needed for molecularly targeted agents where an adult MTD was not defined due to tolerability at doses that met target–effect end points or pharmacokinetic end points prior to observing dose-limiting toxicity. In these agents, the adult MTD could be the initial dose in the pediatric Phase I trial with one or two planned dose escalations.

In situations where new agents for molecular targets are to be tested in combination with standard therapy, a single Phase I study of the targeted agent plus standard therapy can be performed. Where possible, the single agent may be administered alone first in the patients in order to assess toxicity and pharmacokinetics of the targeted agent, followed by use of the agent in combination with standard therapy.

Phase II trials

New agents studied in pediatric cancers have usually been studied previously in adults. A large number of novel agents for diverse molecular targets are under development for adults with cancer, with the promising ones then available for study in pediatric cancers. A challenge is identifying the subset of these targeted agents that could potentially be active in specific pediatric cancer populations. Identification of promising agents for further study is typically based on results from conventional Phase II trials run on patients with refractory/relapsed disease. Improved outcomes for newly diagnosed patients have resulted in fewer patients with relapsed disease, resulting in the availability of few patients for these trials. Hence, more emphasis is being placed on getting robust preclinical data as a source of single-agent information in place of the conventional single-agent Phase I/II trials, with subsequent Phase II trials then being conducted on the combination of novel agents with standard chemotherapy backbones.

The primary objective of Phase II trials is to screen an agent for antitumor activity in order to assess if it should be taken forward for further testing in a Phase III setting. Effective assessments prior to committing to the conduct of a Phase III study minimize the number of patients who are exposed to potentially inferior or highly toxic therapy and may also result in shorter timelines to identifying more effective treatments. Phase II studies also enable us to get a better toxicity profile of the new agent being tested, often in the context of more traditional chemotherapy delivered at recurrence. A positive Phase II trial results in the identification of a single agent or combination therapy that has both promising efficacy and acceptable toxicity.

Phase II trials are typically run in relapsed/refractory patients who may not respond to standard therapy. The primary end point in ALL is usually complete remission (CR) after 4 weeks of treatment. This end point is typically used since it can be observed early in the course of ALL therapy, which tends to be long. Since there are currently a variety of active agents for treating ALL, there is great interest in studying combinations of such drugs. Addition of molecularly targeted agents to standard therapy is of particular interest now since there are often little or no overlapping toxicities between them. In the context of single-arm Phase II trials, when assessing whether the CR rate in patients treated with a combination including an active agent is better than the CR rate for active agent alone, it is difficult to attribute the improved CR rate to the combination only and assume there are no other biases due to patient selection. Randomized, controlled screening designs and selection designs (discussed later) can be used to address these issues.

The Simon two-stage design is a single-arm design for the study of single agents or combinations using relatively small sample sizes [14]. The single-arm design is relatively simple and generally requires a small sample size and short study duration. In order to make an unbiased assessment of the experimental regimen's activity, it is important that the patients used to fix the null hypothesis response rate (that expected with 'standard' therapy) are similar to the patients eligible for the single-arm Phase II trial. An unpromising level of activity ('null hypothesis', p_0) is specified (typically set to 5%), and a promising level of activity ('alternative hypothesis', p_a) is also specified (typically set to 20%). This was historically based on the idea that a minimum CR rate of 20% was needed for patients with recurrent disease, in order to produce a clinically meaningful OS benefit in the subsequent Phase III testing. While compared with recent results the CR rates might seem to be low, it should be noted that second remission rates are very low in poor prognosis subgroups of pediatric ALL, including relapsed/refractory ALL. The minimum number of observed CRs needed to declare the agent to be promising is specified. The sample size and number of CRs are determined such that the type I error rate (probability of concluding an inactive agent is active) is no greater than a specified level α and the type II error rate (the probability of concluding an active agent is inactive) is no greater than a specified level β. A test for futility is performed after a fixed number of initial patients are accrued and evaluated for CR. If the futility criterion is met, indicating that the new agent has a low probability of being declared to be active at the end of the study, the study is stopped early. This ensures that few patients are exposed to a possibly toxic agent with minimal or no activity. With the above design specifications ($p_0 = 0.05$, $p_a = 0.2$, $\alpha = 0.1$ and $\beta = 0.1$), the study would accrue 12 patients in the first stage and if no CRs are observed, the study would close for lack of activity; otherwise, an additional 25 patients would be accrued. If four or more CRs are observed, the drug is declared to be sufficiently active in the disease to justify further study. The development of nelarabine for T-cell ALL illustrates this design. In the pediatric Phase II trial of nelarabine in ALL patients in first relapse, a two-stage design was used with α and β both set at 0.1 and with

a response rate of 35% proving activity (p_a = 0.35) and a response rate of 15% (p_0 = 0.15) indicating inactivity in T-cell ALL.

As discussed earlier, single-agent Phase II trials in ALL are increasingly difficult to conduct. When studying combinations of active agents with standard therapy, it is of interest to determine whether the new agent or combination has a CR rate higher than the CR rate for the standard therapy. The true CR rate for the active agent(s) is unknown and the observed rates from other (typically small) studies where patients may have heterogeneous clinical characteristics have wide variability associated with them. In a single-arm study, misspecifying the null CR rate can have serious implications on the type I and II errors. Consider the addition of a novel agent to standard reinduction therapy for children with an early first relapse of ALL, for whom the standard reinduction therapy is effective in inducing remission in approximately 70% of patients. Table 1 gives the error rates in a single-arm study when the null rate p_0 is either specified too low or too high [15]. The misspecification of the null rate could be due to several reasons. The expected outcomes of historical controls may not accurately represent the expected outcomes of the current patients on experimental therapy. Expected outcomes on the standard regimen may change over time due to changes in supportive care, changes in diagnostic procedures, and differences in prognostic factors between the two patient populations. Consequently, activity of the drug in the current experimental cohort could be different from that seen in the historical cohort. When the true rate is <70% and the agent is truly active, the probability of concluding that the agent is active is lower than the desired level of 0.90, which could result in a promising agent being overlooked. When the true rate is >70%, the probability of concluding the agent is active is higher than the desired level of 10%. This could result in the Phase III trial being conducted with an inactive agent resulting in patients being exposed to an inactive toxic treatment. This was a problem faced in the design of COG trials for relapsed ALL. Patient numbers are too small for a randomized Phase III trial and the historical control data were too old for meaningful comparisons with a new single arm Phase II study. Thus the current approach is to use randomized Phase II designs in this population. The new study in development has concurrent randomization to control (backbone regimen) versus backbone plus novel agent and the objective is to look for preliminary evidence of activity in the relapsed ALL population.

Although it has been proven that tyrosine kinase inhibitors given in combination therapy can significantly improve outcomes in patients, as with imatinib for patients with Ph$^+$ ALL, the problem that remains is that of insufficient patient numbers for a randomized Phase III trial. In addition, any results from this trial may not necessarily be taken forward directly to apply to all frontline ALL, as this is a targeted agent. However, it is worth noting that in the relapsed ALL setting, investigational agents that have shown promise in Phase II trials in this population are now being considered for use in randomized Phase III trials in frontline ALL.

In this era of targeted therapies for new molecular targets, it may not be possible to identify historical controls that expressed the molecular target, on past studies. Thus, investigators may find a randomized Phase II trial with a prospective control more attractive than a single-arm trial with comparisons to a historical control [16–18]. Screening designs enable comparisons of an experimental regimen (standard regimen + novel agent) versus standard regimen alone and provide evidence justifying consideration of an appropriately designed Phase III study in newly diagnosed patients. The randomization minimizes the biases described above. Rubinstein *et al.* proposed a screening design in which the type I error is set to 0.15 or 0.2 and the type II error is set to 0.2 [16]. These error rates are larger than the 0.1 level commonly used in single-arm Phase II trials. Since typical randomized trials require larger numbers of patients, the increased error rates allow for smaller sample sizes for these randomized Phase II trials. Although error rates may seem to be smaller for a single-arm study, in the example given above (Table 1), it can be seen that the type II error rate is larger than 0.2 when the null rate specified is too high by a difference of 2.4% and the type I error rate is larger than 0.2 when the true null rate has been specified too low by a difference of 3%. Therefore, it is quite likely that the error rates for a single-arm study will be larger than the rate of 0.2 specified for the randomized screening design. In the randomized design, the type I error will always be maintained and power for specific alternatives will be known.

Table 1. Probability of concluding that an agent has insufficient or sufficient activity when the null response rate is either specified too low, correctly or too high.

Null response rate	True null rate (%)	Type I error	Type II error†
p_0 is specified too high	60	0.0022	0.689
	65	0.018	0.35
	67.6	0.046	0.20
Correct p_0	70	0.10	0.10
p_0 is specified too low	73	0.20	0.03
	75	0.31	0.01
	80	0.64	<0.01

Assumed null response rate of 70% (p_0) and alternative rate of 85% (p_a); sample size of 59 patients for type I and type II error rates of 0.1 each, respectively.
†The agent increases activity by 15% over the true null rate.

It is important to consider the impact of the results of a positive or negative randomized screening design on the conduct of a subsequent Phase III trial. A positive randomized Phase II study with an EFS end point may make it difficult to then run a well-designed Phase III trial, making it more appropriate to use an early end point like CR rate in the Phase II study and use EFS in the Phase III study. In addition, the Phase II trial may be conducted in relapsed/refractory patients and the definitive Phase III study may then be conducted in patients with newly diagnosed ALL.

Randomized selection designs are useful when the objective is to study and select from combination therapies including a single standard backbone regimen with the addition of several experimental agents [19]. The experimental agents chosen for study in a selection design must have data from previous trials showing some activity and comparable toxicity profiles. Where appropriate, the randomized Phase II selection design can be implemented with modest sample sizes. Simon *et al.* demonstrate that only 29–37 patients per arm will yield 90% power to detect a regimen that has a response rate superior by 15%, in a two-armed study. In this design, unlike a Phase III setting, there is no formal comparison of arms. The arm with the best activity (e.g., CR rate) is then chosen to take forward to a Phase III trial. A drawback of this design is that the 'best' arm chosen may have an observed CR rate that is very low. To prevent this, an appropriate futility rule using the standard two-stage single-arm Phase II design, can be incorporated into each arm to ensure a minimum level of activity. Jung proposed a two-stage comparative design allowing for early termination of the study when the experimental arm does not show promising efficacy over the prospective control at interim analysis [20]. Randomized selection designs minimize biases introduced in evaluating multiple combination regimens, due to selection bias, changes in evaluation criteria, patient care, or supportive care across participating centers. While selection designs do provide much stronger comparative data than a series of single-arm trials to select promising treatment(s), the lack of direct comparison to a prospective control could make the investigators less confident in the trial outcomes. In addition, if there is high confidence in the historical data it would be more efficient to not randomize and instead run a single-arm trial with the experimental regimen, using half the number of patients and comparisons can be made to the historical controls.

Phase III trials

Most new treatments in ALL produce either no benefit or a moderate improvement in outcomes, with large benefits being observed very rarely. Nevertheless, it is important to confirm this in the setting of a randomized-controled trial (RCT) where the new treatment is compared with the standard treatment in use. If a nonrandomized study is performed using historical controls, other factors in addition to the introduction of a new treatment may have changed over time. This includes changes in patient care, supportive care, diagnostic methods, disease classification and staging, and other presenting characteristics in the patients, all of which could influence outcomes for the new treatment. This could result in a possibly ineffective treatment, with serious toxicity and costs being adopted as standard of care based on the results of a poorly designed single-arm study.

Randomization ensures that biases introduced due to confounding factors are minimized; distribution of prognostic factors is balanced on the two arms, and the patient cohorts on the two arms are similar with respect to presenting clinical characteristics. The process of randomization ensures that the groups being compared, on average, are similar in all respects other than the treatments being evaluated, with any differences being due to chance.

Multiple RCTs in pediatric ALL have been successful in identifying more effective treatments over the years that have defined current standard of care in ALL. Examples of these trials include:

- Augmented Berlin-Frankfurt-Munster (BFM) therapy consisting of longer and stronger postinduction intensification, was shown on CCG 1882 to be superior for high-risk patients with slow, early (day 7) response to induction with better 5-year EFS compared with standard therapy (72.6% ± 3.9% vs 57.0 ± 4.2%; p = 0.0008) [21].

- The superiority of dexamethasone in induction and maintenance compared with prednisone for standard risk ALL was proven on CCG-1922; 6-year EFS of 85.5% ± 1.7% in patients randomized to dexamethasone and 79.1% ± 1.9% in patients randomized to prednisone (p = 0.002) [22]. This advantage was also supported by results of recent Medical Research Council and BFM trials [23,24].

- CCG-1961 showed that stronger but not longer postinduction intensification (augmented BFM regimen without second delayed intensification) is more beneficial for high-risk patients with a rapid, early (day 7) response to induction therapy with 5-year EFS 82.2 ± 1.6% for stronger versus 72.5 ± 1.9% for standard postinduction therapy, respectively (p = 0.0001) [25].

- CCG-1991 established that escalating intravenous methotrexate without leucovorin rescue is superior to oral methotrexate during interim maintenance in standard-risk ALL; 5-year EFS 92.6 ± 1.2% versus 88.7 ± 1.4%, respectively (p = 0.009) [26].

- COG AALL0232 established the superiority of high-dose methotrexate over Capizzi methotrexate in high-risk ALL; 5-year EFS 82.0 ± 3.4% and 75.4 ± 3.6%, respectively (p = 0.006) [27].

While OS provides the absolute result as to whether a new treatment results in fewer deaths, EFS is usually the preferred primary end point in pediatric cancer trials. Use of OS as an end point may not be practical for evaluating therapies in ALL where OS rates are very high (90–95%), and death from disease may occur a long time after start of therapy and other events (e.g., allogeneic transplantation after relapse) may occur in the interim, influencing the result of the trial. Variation in subsequent therapies delivered following first recurrence can have an effect on OS, the primary end point for the first study. Since the event rate is higher and the rate of long-term survival following recurrent disease is low, use of EFS gives better power and hence lower sample sizes to detect a difference. Due to lack of evidence supporting their predictive value, early surrogate end points including end induction blast count and MRD, are less useful as end points for RCTs despite being highly prognostic of outcome.

Clinical outcomes on new treatments need to be balanced with related toxicities, for acute, short-term and late effects (e.g., infertility and intellectual impairment). As outcomes improve and patients live longer, health-related quality of life needs to be given more importance, and justifies the incorporation of health-related quality of life secondary end points into clinical trials. However, in the pediatric setting, occurrence of short-term toxicities (even 'life-threatening' ones) that resolve quickly is acceptable if it then results in better long-term survival rates together with high quality of survivorship. This is critical when considering reduction in therapy questions for patient subgroups identified to have a very low risk of relapse, and the risk–benefit of lower toxicities versus increased risk of relapse should be carefully considered.

Conventional sample size calculations for RCTs have been based on standard choices for type I error (α) and power. A two-sided 0.05 level test is commonly used to compare treatment arms in RCTs. Alternatively, a one-sided test may be used if there is no interest to prove that a new treatment is worse than the standard regimen. In this case, a significance level of 0.025 is frequently used in adult cancer clinical trials, but with the smaller number of pediatric patients use of a one-sided α of 0.05 is considered acceptable to provide convincing evidence of efficacy. Power for the trials is typically set at 80–90%, requiring large sample sizes in order to detect the expected moderate improvements in outcome. In poorer prognosis subsets of patients, the numbers needed will be smaller whereas a larger number of patients are needed in trials involving good prognosis patients.

Designs for rare ALL subsets

When designing trials with rare subsets of ALL (including infant ALL, Ph$^+$ ALL, JAK mutation-positive ALL, or Down syndrome ALL), the smaller patient numbers may make design of an RCT in that disease impossible. Due to the lack of power in such trials, frequently non-randomized trials with historical comparison may be undertaken. An alternative is to design a randomized trial based on conventional sample size calculations, but using a higher α (e.g., 0.2) or lower power (e.g., 0.7) [28]. Consider the example of infants with mixed-lineage leukemia (MLL) rearranged ALL (MLL-R) who have very poor outcomes with 3-year EFS rates below 50%. The current COG trial for infant ALL, tests efficacy of the FLT3 inhibitor lestaurtinib (a strategy with a strong biological rationale) in MLL-R patients. MLL-R patients are randomized to ± lestaurtinib in combination with a common backbone therapy at the end of induction. Designing a two-arm randomized trial with the usual parameters, is not feasible in this population due to the low accrual rate (50 patients/year). The study was designed using a one-sided relaxed type I error rate of 15%, and 80% power to look for an improvement in 3-year EFS from 50 to 65%.

As outcomes for ALL improve, there is interest in reduction of therapy in specific low-risk subsets in order to determine if good outcomes can be maintained using therapies with fewer short- and long-term toxicities. Sample size consideration for equivalence or noninferiority trials are different from those for superiority trials described above [29]. Sample sizes are larger for equivalence trials even with the use of one-sided tests, requiring use of large margins to prove equivalence, which may not be clinically acceptable. The alternatives suggested for superiority trials can then be applied to the design of noninferiority trials (e.g., looking at therapy reduction questions for subgroups at low risk for relapse) higher (e.g., 90%) power, with a higher type I error rate (e.g., 10%) [30,31].

Incidence of childhood ALL is high enough where use of RCTs is feasible in most large subgroups, including National Cancer Institute standard- and high-risk patients (e.g., 1200 and 600 patients per year, respectively, in COG). However, in rare subtypes – including infants, Ph$^+$ ALL and relapsed ALL – patient numbers are too small to allow for the use of conventionally designed RCTs. Alternative strategies in such cases,

could include international collaborations with other groups in designing and running intergroup trials. Running such trials can be a challenge from getting consensus on study questions and design, to meeting regulatory requirements in different countries. Some of the logistical issues can be resolved with the use of electronic/web-based data entry systems. COG collaborated with the European Study Group for Philadelphia Chromosome-Positive ALL group to develop a transatlantic pediatric Ph$^+$ ALL trial (AALL1122) of tyrosine kinase inhibitor dasatinib plus chemotherapy which is currently active and enrolling patients.

Single-arm trials

Use of RCTs in evaluating the efficacy of molecularly targeted new agents in small biologic subgroups may be challenging, even with the alternatives suggested above. Single-arm trials with comparison with historical controls have been used in this setting when data for controls on the standard therapy are available or when outcomes on past trials have been very poor for these subgroups. For example, the COG trial AALL0031 for Ph$^+$ ALL was designed as a single-arm trial with imatinib given in combination with an intensive chemotherapy backbone. Besides early end points looking at MRD before and after imatinib administration in the first two blocks of consolidation therapy, EFS was to be compared with historical controls from prior COG trials. Outcome in these patients on past trials was very poor (3-year EFS: 35%) compared with that observed when imatinib was added to standard therapy on AALL0031 (3-year EFS: 80%; $p < 0.0001$) providing stronge evidence supporting the efficacy of imatinib + backbone therapy for Ph$^+$ ALL. However, such large improvements in outcomes are not common. With more moderate expected improvements, it is still beneficial to consider randomized trials with adjusted type I and II error rates.

Another alternative approach to RCTs that has been used in other pediatric malignancies is to run a single-arm study where the new therapy is compared with some fixed standard. One defines a time-to-event distribution based on the historical experience and assumes this is the null hypothesis distribution for the new treatment. The outcome for the new treatment is then compared with this fixed standard. This approach differs from historically controlled studies, in that the randomness of the outcome for the comparison group has been removed. For example, the historical EFS experience for patients with Ph$^+$ ALL as reported in Schultz *et al.* [9], can be described reasonably well for the first 3 years of follow up by the EFS function:

$$s(t) = 0.30 + 0.7 \times \exp(-t) \quad \text{Equation 1}$$

With 1-, 2- and 5-year EFS values of 55, 40 and 30%, respectively.

Using such an approach in the context of the COG is reasonable for the following reasons. First, a large historical experience often exists; that is, the estimated values typically have small variability. Second, the institutional membership of the COG has been largely stable and enrollment rates consistent; that is, we see a stable pipeline of study enrollments, suggesting that there is little variability of patient characteristics across sequential studies.

The development of statistical considerations, including the calculation of the required sample size, is straightforward; the details can be found in Finkelstein *et al.* [32]. Testing for an improvement or a reduction in outcome and interim monitoring of the emerging results is also straightforward, using the one-sample log-rank test of Woolson [33]. The test statistic compares the number of observed events ('observed', O) with the number events expected if the time-to-event distribution was that specified by the null hypothesis, even the observed follow up ('expected', E). The test statistic is $T = (O - E)^2/E$, which has a large sample chi-square distribution with one degree of freedom.

A COG study that employed this technique was D9602, a study of the treatment of low-risk rhabdomyosarcoma [34]. The historical experience in treating such patients suggested that the long-term EFS with standard treatment would be 85%, with most events observed by 3 years of follow up. The available patients for study preclude the conduct of a randomized study. However, there was interest in determining whether standard therapy could be revised to reduce short- and long-term adverse events without reducing long-term EFS, The sample size (264 eligible patients) was chosen so that there was expected to be 90% power to detect a decrease in the long-term EFS to 78% (a 50% increase in the failure rate).

This study design has also been used for studies of subsets of patients with neuroblastoma or Wilms tumor. It is best used in situations where the number of available patients to study and their expected outcome precludes the conduct of a properly powered randomized study and, there can be agreement on a fixed time-to-event outcome that is expected for patients under study and the goal is to demonstrate improvement in outcome or exclude a deficit in outcome with a new therapy.

Factorial designs

Factorial designs are an efficient way of addressing more than one primary question of interest in a clinical trial. COG Phase III trials for ALL frequently use 2 × 2 factorial designs, allowing two questions to be

answered with about the same patient resources that would be necessary to answer one question. It is, of course, critical to ensure that the treatments being evaluated are compatible and can be given together on the same trial. Factorial designs may involve allocation to more than one randomization at the same time (e.g., COG AALL0232 had one upfront randomization to prednisone vs dexamethasone in induction, and a later randomization in interim maintenance to high-dose methotrexate vs Capizzi methotrexate). Sample size for a factorial design is similar to that for a single question trial, if there is no interaction between the two factors/treatments. Presence of negative interaction between treatments (i.e., the benefit of the combination is less than the sum of the individual benefits), the sample size has to be increased, but is still less than that needed for two separate trials to be conducted. Factorial designs are recommended only when lack of qualitative interaction between the two interventions can be assumed. In the absence of interaction, stratified analysis can be used for estimation of main effects. In the presence of significant interaction, various subset analyses may be conducted to compare individual treatments, keeping in mind that the study was not powered for such comparisons [35].

Future perspective

Randomized trials provide a robust means of evaluation of new treatments for pediatric ALL. These trials have helped identify both effective and ineffective interventions, resulting in improved outcomes for patients. Design and conduct of clinical trials in ALL will have numerous challenges in the future, including choice of treatment approaches to be studied in the various ALL subsets and the problem of small numbers of patients to study as we further subdivide patients into subsets based on presence of specific biomarkers and molecular targets. The goals for the future are to better understand the molecular pathways leading to specific phenotypes, to minimize the risk of relapse by identifying subsets of patients requiring more intensive therapy and to minimize the risk of adverse events for those patients highly likely to be cured with currently available therapies.

Disclaimer

Opinions expressed in this article are those of the authors alone and do not represent those of any medical society or professional organization.

Financial & competing interests disclosure

This work was supported by the Chair's grant U10 CA98543, Statistics and Data Center grant U10 CA98413 of the Children's Oncology Group from the National Cancer Institute, NIH, MD, USA. A complete listing of grant support for research conducted by the Children's Cancer Group and Pediatric Oncology Group before initiation of the COG grant in 2003 is available online at: www.childrensoncologygroup.org/admin/grantinfo.htm. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

- Although outcomes have improved significantly over the years for pediatric acute lymphoblastic leukemia (ALL), there are many clinically and biologically defined subsets with very poor outcomes. Some of these rare subsets have very low annual accrual numbers making traditional randomized Phase III trials infeasible. This has been the primary motivation for the development of innovative designs for effectively testing targeted agents in these subgroups.

Phase I trials

- Phase I trials are used (usually in the relapsed ALL population) to determine the maximum-tolerated dose and assess pharmacokinetics of new agents either administered singly or in combination with other standard drugs.

Phase II trials

- Phase II trials are traditionally used to screen new agents given alone or in combination for antitumor activity, and determine if any should be taken forward into Phase III testing. Single-arm Phase II trials used historically, have several issues associated with their design, which can be rectified with the use of, when appropriate, randomized Phase II trials.

Phase III trials

- Phase III trials are the gold standard for establishing the efficacy of an experimental regimen compared with standard of care. However, it is infeasible to run them in rare subsets of ALL with low accrual numbers. This has motivated the development of innovative trial designs that can be used for effective testing of a targeted agent alone or in combination in these subgroups.

Future perspective

- Identification of further smaller subgroups in the ALL population based on presence of specific biomarkers and molecular targets, will raise numerous challenges in designing trials for the efficient assessment of potentially effective therapies in these groups. In order to achieve future goals, more effort needs to be invested in the area of innovative trial design.

References

Papers of special note have been highlighted as:

- ■ of interest
- ■■ of considerable interest

1 Ries LA, Smith MA, Gurney JG *et al.* Cancer incidence and survival among children and adolescents: United States Seer Program 1975–1995. Bethesda, MD, USA (1999).

2 Smith MA, Seibel NL, Altekruse SF *et al.* Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *J. Clin. Oncol.* 28, 2625–2634 (2010).

3 Hunger SP, Lu X, Devidas M *et al.* Improved survival for children and adolescents with acute lymphoblastic leukemia between 1990 and 2005: a report from the children's oncology group. *J. Clin.Oncol.* 30(14), 1663–1669 (2012).

4 Hunger SP, Loh ML, Whitlock JA *et al.* Children's Oncology Group's 2013 blueprint for research: acute lymphoblastic leukemia. on behalf of the COG Acute Lymphoblastic Leukemia Committee. *Pediatr. Blood Cancer* 60(6), 957–963 (2013).

■■ **Gives a complete summary of recent Children's Oncology Group research accomplishments in pediatric acute lymphoblastic leukemia (ALL).**

5 Pieters R, Carroll WL. Biology and treatment of acute lymphoblastic leukemia. *Pediatr. Clin. North Am.* 55(1), 1–20 (2008).

6 Schultz KR, Pullen DJ, Sather HN *et al.* Risk-and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood* 109, 926–935 (2007).

7 Borowitz MJ, Devidas M, Hunger SP *et al.* Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a children's oncology group study. *Blood* 111, 5477–5485 (2008).

■ **Provides evidence of the strong prognostic influence of minimal residual disease in ALL. Minimal residual disease end of induction therapy has been on later studies confirmed to be the most important prognostic factor in ALL.**

8 Hunger SP, Raetz EA, Loh ML *et al.* Improving outcomes for high-risk ALL: translating new discoveries into clinical care. *Pediatr. Blood Cancer* 56, 984–993 (2011).

9 Schultz KR, Bowman WP, Aledo A *et al.* Improved early event-free survival with imatinib in philadelphia chromosome-positive acute lymphoblastic leukemia: a Children's Oncology Group study. *J. Clin. Oncol.* 27, 5175–5181 (2009).

■■ **Report on the outcomes from the first pediatric trial for Philidelphia-positive ALL using a tyrosine kinase inhibitor, run by the Children's Oncology Group. It had impressive early event-free survival rates of 80% compared with historical rates around 30%.**

10 Mullighan CG, Zhang J, Harvey RC *et al.* JAK mutations in high-risk childhood acute lymphoblastic leukemia. *Proc. Natl Acad. Sci. USA* 106(23), 9414–9418 (2009).

11 Smith M, Bernstein M, Bleyer WA *et al.* Conduct of Phase I trials in children with cancer. *J. Clin. Oncol.* 16(3), 966–978 (1998).

12 Skolnik JM, Barrett JS, Jayaraman B *et al.* Shortening the timeline of pediatric Phase I trials: the rolling six design. *J. Clin. Oncol.* 26(2), 190–195 (2008).

13 Onar A, Kocak M, Boyett JM. Continual reassessment method vs traditional empirically based design: modifications motivated by Phase I trials in pediatric oncology by the Pediatric Brain Tumor Consortium. *J. Biopharm. Stat.* 19(3), 437–455 (2009).

14 Simon, R. Optimal two-stage designs for Phase II clinical trials. *Control Clin Trials* 10(1), 1–10 (1989).

15 Smith M, Devidas M, Wheatley K *et al.* Strategies for new agent development and clinical trial considerations. In: *Childhood Leukemia: A Practical Handbook.* Reaman GH, Smith FO (Eds). Springer-Verlag, Berlin Heidelberg, Germany, 215–244 (2011).

16 Rubinstein LV, Korn EL, Freidlin B *et al.* Design issues of randomized Phase II trials and a proposal for Phase II screening trials. *J. Clin. Oncol.* 23(28), 7199–7206 (2005).

17 Ratain MJ, Sargent DJ. Optimizing the design of Phase II oncology trials: the importance of randomization. *Eur. J. Cancer* 45, 275–280 (2009).

18 Rubinstein L, Crowley J, Ivy P *et al.* Randomized Phase II designs. *Clin. Cancer Res.* 15, 1883–1890 (2009).

19 Simon R, Wittes RE, Ellenberg SS. Randomized Phase II clinical trials. *Cancer Treat. Rep.* 69(12), 1375–1381 (1985).

20 Jung SH. Randomized Phase II trials with a prospective control. *Stat. Med.* 27, 568–583 (2008).

21 Nachman JB, Sather HN, Sensel MG *et al.* Augmented post-induction therapy for children with high-risk acute lymphoblastic leukemia and a slow response to initial therapy. *N. Engl. J. Med.* 338(23), 1663–1671 (1998).

22 Bostrom BC, Sensel MR, Sather HN *et al.* Dexamethasone versus prednisone and daily oral versus weekly intravenous mercaptopurine for patients with standard-risk acute lymphoblastic leukemia: a report from the Children's Cancer Group. *Blood* 101(10), 3809–3817 (2003).

23 Mitchell CD, Richards SM, Kinsey SE *et al.* Benefit of dexamethasone compared with prednisolone for childhood acute lymphoblastic leukaemia: results of the UK Medical Research Council ALL97 randomized trial. *Br. J. Haematol.* 129(6), 734–745 (2005).

24 Schrappe M, Zimmermann M, Moricke A *et al.* Dexamethasone in induction can eliminate one third of all relapses in childhood acute lymphoblastic leukemia (ALL): results of an international randomized trial in 3655 patients (Trial AIEOP-BFM ALL 2000). Presented at: *50th Annual Meeting of the American Society of Hematology.* San Francisco, CA, USA, 6–9 December 2008.

25 Seibel NL, Steinherz PG, Sather HN *et al.* Early postinduction intensification therapy improves survival for children and adolescents with high-risk acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood* 111(5), 2548–2555 (2008).

26 Matloub Y, Bostrom BC, Hunger SP *et al.* Escalating intravenous methotrexate improves event-free survival in children with standard-risk acute lymphoblastic leukemia: a report for the children's oncology group. *Blood* (24)118, 243–251 (2011).

27 Larsen EC, Salzer WL, Devidas M *et al.* High dose methotrexate (Hd-mtx) as compared with Capizzi methotrexate plus asparaginase (C-mtx/asnase) improves event-free survival (EFS) in children and young adults with high-risk acute lymphoblastic leukemia: a report from the Children's Oncology Group study AALL0232. Presented at: *Annual Meeting of American Society of Clinical Oncology.* Chicago, IL, USA, 3–7 June 2011.

28 Sposto R, Stram DO. A strategic view of randomized trial design in low-incidence paediatric cancer. *Stat. Med.* 18(10), 1183–1197 (1999).

29 Piaggio G, Elbourne DR, Pocock SJ *et al.* Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 295(10), 1152–1160 (2006).

30 Chan IS. Power and sample size determination for noninferiority trials using an exact method. *J. Biopharm. Stat.* 12(4), 457–469 (2002).

31 Wang SJ, Hung HM. Assessing treatment efficacy in noninferiority trials. *Control Clin. Trials.* 24(2), 147–155 (2003).

32 Finkelstein DM, Muzikansky A, Schoenfeld DA. Comparing survival of a sample to that of a standard population. *J. Natl Cancer Inst.* 95, 1434–1439 (2003).

33 Woolson RF. Rank tests and a one-sample logrank test for comparing observed survival data to a standard population. *Biometrics* 37, 687–696 (1981).

34 Raney RB, Walterhouse DO, Meza JL *et al.* Results of the Intergroup Rhabdomyosarcoma Study Group D9602 protocol, using vincristine and dactinomycin with or without cyclophosphamide and radiation therapy, for newly diagnosed patients with low-risk embryonal rhabdomyosarcoma: a report from the Soft Tissue Sarcoma Committee of the Children's Oncology Group. *J. Clin. Oncol.* 29, 1312–1318 (2011).

35 Green S. Factorial designs with time-to-event end points. In: *Handbook of Statistics in Clinical Oncology*. Crowley J (Ed.). Marcel Dekker, New York, NY, USA, 161–171 (2001).