Published in final edited form as:

Contemp Clin Trials. 2009 May ; 30(3): 279–281. doi:10.1016/j.cct.2009.02.002.

Treatment-Subgroup Interaction: An Example from a Published, Phase II Clinical Trial

Carolyn E. Behrendt, PhD¹ and Edmund A. Gehan, PhD²

¹Department of Biostatistics, City of Hope National Medical Center and Beckman Research Institute, Duarte, CA, USA

²Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University School of Medicine, Washington, DC, USA

Abstract

Phase II trial designs that ignore between-patient heterogeneity and do not allow for treatment-subgroup interactions may produce very large false positive and false negative error rates if efficacy varies by subgroup. Recent discussions of this problem were illustrated with scenarios and computer simulations. In this short communication, we reanalyzed a published phase II trial to highlight the need to consider between-patient heterogeneity and the possibility of treatment-subgroup interaction when designing and analyzing phase II studies. The single-arm trial evaluated amsacrine plus cytosine arabinoside, vincristine, and prednisone (a combination abbreviated as OAP) for adult acute leukemia, when standard treatment was adriamycin plus OAP. We carried out an analysis of covariance (ANCOVA) incorporating data from historical control patients who met eligibility criteria for the trial and received standard treatment at the study center in the years immediately preceding the trial. Patients administered experimental treatment and control patients were classified as having favorable or unfavorable prognosis according to their predicted probability of response to standard treatment. When the prognostic subgroup of patients was ignored, the response rates for experimental and standard treatment appeared similar. However, fitting an ANCOVA model determined that the effects of subgroup, treatment, and their interaction were statistically significant: experimental treatment was superior to standard treatment in patients with unfavorable prognosis and inferior to standard treatment in patients with favorable prognosis. This real-world example of treatment-subgroup interaction highlights the need to employ phase II designs that consider between-patient heterogeneity and the possibility that efficacy differs by subgroup.

Background

Nearly all designs for phase II trials make the simplifying assumption that patients are alike in their probability of responding to experimental treatment [1], [2] and [3]. Often this assumption is not valid, and the patients entering the trial belong to subgroups with differing probabilities of response. When this occurs, phase II designs that ignore the possibility that efficacy differs by subgroup may produce extremely large false positive and false negative error rates within subgroups [4]. Recent discussions of this problem were illustrated with scenarios and computer simulations [4] and [5]. In this short communication, we reanalyzed a published phase II trial

Carolyn E. Behrendt, PhD, City of Hope, 1500 E. Duarte Road, Duarte, CA 91010, USA, e-mail: cbehrendt@coh.org, tel: 626-256-4673, ext 61148, fax: 626-471-7106.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

[6] to highlight the need to consider between-patient heterogeneity and the possibility of treatment-subgroup interaction when designing and analyzing phase II studies.

Methods

The single-arm trial that we reanalyzed evaluated amsacrine (AMSA) plus cytosine arabinoside, vincristine, and prednisone (a combination abbreviated as OAP) for adult acute leukemia, when standard treatment was adriamycin plus OAP [6]. For ethical reasons, experimental treatment was initially restricted to patients with unfavorable prognosis on standard treatment; as results on the trial accumulated, the criteria for administering experimental treatment were broadened until ultimately all patients received experimental treatment.

Subjects who received experimental treatment ($n=134$) were classified into two subgroups based on predicted probability of response (PPR) to standard treatment, calculated from a logistic regression model constructed based on 300 patients [7] and validated using another 107 patients [8], all of whom met eligibility criteria for the trial and were treated at the same institution in the years immediately preceding the trial. Patients in the favorable prognostic subgroup had a PPR of at least 0.60.

Data for the current reanalysis were obtained from tables in publications reporting the trial [6] and the PPR model [7] and [8]. The 407 patients described above [7] and [8] who received standard treatment served as our historical control patients and were classified into two prognostic subgroups in the same manner as the subjects who received experimental treatment. We carried out an analysis of covariance (ANCOVA) where response was the number of complete remissions (CR) out of total subjects treated (N), or CR/N.

Results

Response to treatment was as shown in Table 1. When the prognostic subgroup of patients was ignored, the response rates for experimental and standard treatment did not differ significantly (52.2% versus 60.2%, chi-square test, $p>0.10$). However, fitting an ANCOVA model determined that the effects of subgroup, treatment, and their interaction were statistically significant (Table 2). Thus experimental treatment was superior to standard treatment in patients with unfavorable prognosis and inferior to standard treatment in patients with favorable prognosis (Table 3).

Discussion

The current reanalysis of a phase II trial [6] detected significant treatment-subgroup interaction. This real-world example highlights the need for single-arm trial designs to take into account between-patient heterogeneity and allow for the possibility of differing treatment efficacy among subgroups.

Our reanalysis of the AMSA trial incorporated data from appropriate historical control patients [10] into an ANCOVA. Investigators planning future phase II trials might consider employing the Bayesian design recently proposed by Wathen et al. [5], in which historical subgroup effects of standard treatment are incorporated into an ANCOVA as informative priors. Because the design allows for treatment-subgroup interaction and subgroup-specific stopping rules, accrual may be stopped within one subgroup but continue in another.

However, it should be noted that the design of Wathen et al. [5] would not have been appropriate for the AMSA trial, in which subjects were previously untreated patients for whom moderately effective standard treatment was available. For this reason, it would not have been ethical to

administer experimental treatment simultaneously to those with favorable and unfavorable prognosis. Instead, experimental treatment was initially restricted to patients who had the most unfavorable prognosis on standard treatment. Only after their response rate was observed to be higher than predicted were criteria for administering experimental treatment broadened, until ultimately they included all patients entering the trial [6].

The AMSA trial also evaluated toxicity (days of fever, episodes of infection, hyperbilirubinemia, and renal toxicity). According to the report of the trial, experimental treatment was associated with increased frequency of hyperbilirubinemia and fewer days of fever compared to standard treatment [6]. If predicted risks of specific toxicities on standard treatment had been calculated similarly to PPR, subjects receiving experimental treatment and control patients could have been assigned to subgroups of toxicity risk, and ANCOVA models of specific toxicities could have been constructed. In this way, the study could have ascertained whether risk of toxicity, like treatment benefit, differed by prognostic subgroup. The resulting information could have made possible a subgroup-specific risk-benefit assessment of experimental treatment. Unfortunately, published data relating to the trial were not sufficiently detailed to permit us to perform such an analysis, which should be investigated in future trials.

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989;10:1–10. [PubMed: 2702835]
2. Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemp Clin Trials* 2008;29:32–41. [PubMed: 17544337]
3. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med* 1995;14:357–379. [PubMed: 7746977]
4. Thall PF, Wathen JK. Bayesian designs to account for patient heterogeneity in phase II clinical trials. *Curr Opin Oncol* 2008;20:407–411. [PubMed: 18525336]
5. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Stat Med* 2008;27:2802–2815. [PubMed: 17948869]
6. Keating MJ, Gehan EA, Smith TL, et al. A strategy for evaluation of new treatments in untreated patients: application to a clinical trial of AMSA for acute leukemia. *J Clin Oncol* 1987;5:710–721. [PubMed: 3553435]
7. Keating MJ, Smith TL, Gehan EA, et al. A prognostic factor analysis for use in development of predictive models for response in adult acute leukemia. *Cancer* 1982;50:457–465. [PubMed: 7093887]
8. Smith TL, Gehan EA, Keating MJ, Freireich EJ. Prediction of remission in adult acute leukemia: development and testing of predictive models. *Cancer* 1982;50:466–472. [PubMed: 7093888]
9. Hosmer, DW., Jr; Lemeshow, S. *Applied Logistic Regression*. Vol. 2nd. John Wiley & Sons, Inc.; New York: 2000.
10. Gehan EA. The evaluation of therapies: historical control studies. *Stat Med* 1984;3:315–324. [PubMed: 6528131]

Table 1

Complete Remissions (CR) and Total Subjects (N), By Treatment and Prognosis

Prognosis	Trial Subjects, Experimental Treatment (N=134)		Historical Control Patients, Standard Treatment (N=407)	
	CR / N	(%CR)	CR / N	(%CR)
Unfavorable	42 / 91	(46.2%)	57 / 172	(33.1%)
Favorable	28 / 43	(65.1%)	188 / 235	(80.0%)
Total	70 / 134	(52.2%)	245 / 407	(60.2%)

Table 2

Logistic Regression Model of Complete Remission.

	Estimate (SE)	<i>p</i>
Intercept	1.39 (0.16)	<0.0001
Unfavorable Prognosis	-2.09 (0.23)	<0.0001
Experimental Treatment	-0.76 (0.36)	0.0338
Interaction	1.31 (0.45)	0.0034

The Hosmer-Lemeshow goodness-of-fit statistic [9] for the model is zero ($p=1.00$), indicating that the numbers of CR and non-CR responses predicted by the model corresponded exactly to the numbers observed.

Table 3
Relative Odds of Complete Remission with Experimental (N=134) versus Standard Treatment (N=407), Stratified by Prognosis.

Parameter	Odds Ratio (95% CI) ^a	Odds Ratio (95% C I) ^b
Unfavorable Prognosis		
Experimental Treatment	1.73 (1.03, 2.91)	0.21 (0.09, 0.51)
Standard Treatment	1.00	0.12 (0.08, 0.19)
Favorable Prognosis		
Experimental Treatment	3.77 (1.57, 9.04)	0.47 (0.23, 0.94)
Standard Treatment	8.07 (5.14, 12.66)	1.00

^a when the reference category is standard treatment and unfavorable prognosis

^b when the reference category is standard treatment and favorable prognosis