ELSEVIER

Journal of Clinical Epidemiology 60 (2007) 1234−1238

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A simple sample size formula for analysis of covariance in randomized clinical trials

George F. Borm[a,*], Jaap Fransen[b], Wim A.J.G. Lemmens[a]

[a]*Department of Epidemiology and Biostatistics, Radboud University Nijmegen Medical Centre, Geert Grooteplein 21, PO Box 9101, NL-6500 HB Nijmegen, The Netherlands*

[b]*Department of Rheumatology, Radboud University Nijmegen Medical Centre, Geert Grooteplein 21, PO Box 9101, NL-6500 HB Nijmegen, The Netherlands*

Accepted 15 February 2007

Abstract

Objective: Randomized clinical trials that compare two treatments on a continuous outcome can be analyzed using analysis of covariance (ANCOVA) or a *t*-test approach. We present a method for the sample size calculation when ANCOVA is used.

Study Design and Setting: We derived an approximate sample size formula. Simulations were used to verify the accuracy of the formula and to improve the approximation for small trials. The sample size calculations are illustrated in a clinical trial in rheumatoid arthritis.

Results: If the correlation between the outcome measured at baseline and at follow-up is ρ, ANCOVA comparing groups of $(1-\rho^2)n$ subjects has the same power as *t*-test comparing groups of n subjects. When on the same data, ANCOVA is used instead of *t*-test, the precision of the treatment estimate is increased, and the length of the confidence interval is reduced by a factor $\sqrt{1-\rho^2}$.

Conclusion: ANCOVA may considerably reduce the number of patients required for a trial.

Keywords: Power; Sample size; Precision; Analysis of covariance; Clinical trial; Statistical test

1. Introduction

Randomized clinical trials (RCTs) that compare treatment A with treatment B on a continuous outcome measure can be analyzed in several ways. A straightforward option is to compare the follow-up scores at the end of the treatment period (Y_1) using a *t*-test or analysis of variance (ANOVA). When the outcome is also measured at baseline (Y_0), the change scores ($Y_1 - Y_0$) between the treatment groups can be compared, again using a *t*-test. Another approach is to use analysis of covariance (ANCOVA) and to analyze Y_1 or $Y_1 - Y_0$ in a linear regression model that includes treatment group and Y_0 as independent covariates ($Y_1|Y_0$ or $Y_1 - Y_0|Y_0$).

An advantage of the use of ANCOVA is that it adjusts for baseline differences between the treatment groups. ANCOVA also has more statistical power than the *t*-test, so sample size requirements are lower [1−3]. Although this is commonly known, to our knowledge, simple methods for the sample size calculation for ANCOVA have not been available so far. Consequently, when ANCOVA is planned for a trial, this is usually not taken into account in the determination of the sample size, leading to unnecessarily large trials.

We propose a two-step method for the sample size calculation. First, the sample size is calculated as if a *t*-test on the follow-up scores were carried out, then the number of subjects is multiplied by a "design factor" to produce the number of subjects required for the ANCOVA. As the power of an ANCOVA with dependent variable $Y_1 - Y_0$ is the same as the power of an ANCOVA with variable Y_1, we only discuss the latter method.

2. Methods

We assumed that Y_0 and Y_1 were the baseline and outcome variables, respectively, of a clinical trial with two treatment groups. The standard deviation (SD) and the correlation between Y_0 and Y_1 were known. We then calculated the conditional variance of $Y_1|Y_0$. Based on this result we determined the design factor, that is, the ratio between the number of subjects required for an ANCOVA and the number of subjects required for a *t*-test. In practice, the SD and correlation are (implicitly) estimated as a part of

* Corresponding author. Tel.: +31-24-361-7667; fax: +31-24-361-3505.

E-mail address: G.Borm@epib.umcn.nl (G.F. Borm).

0895-4356/07/$ − see front matter © 2007 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2007.02.006

the ANCOVA procedure. This means that the sample size based on known SDs and known correlation coefficient is only approximate. We therefore conducted a simulation study to verify the accuracy of our sample size calculation method. As the method consists of two steps, we evaluated the accuracy of both steps.

We only show results for trials with treatment groups of equal size, but the results for unequal group sizes are similar.

2.1. Accuracy of sample size calculation methods for the t-test

To evaluate the first step, that is, the determination of the sample size required for the *t*-test, we assumed that the treatment differences between the groups ranged from 0.5 SDs up to 1.5 SDs. For each difference, we calculated the sample size that is required for 80% or 90% power, using the sample size calculation package Nquery Advisor® or the formula $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/(m_B - m_A)^2$ [4]. For each combination of treatment difference and group size, we ran 6,400 trials and estimated the "true power" by calculating the percentage of *t*-tests that were statistically significant.

2.2. Accuracy of the sample size calculation method for ANCOVA

The second step of our sample size method is to multiply the sample size for the *t*-test by the design factor. We verified the accuracy of the design factor by simulating clinical trials with baseline and outcome variables Y_0 and Y_1, respectively. The differences between the means of Y_1 in the groups were chosen in such a way that the required sample sizes as calculated using our method were 10, 25, 50, and 100 per treatment arm. The means of Y_0 in the two groups were equal, and the correlation between Y_0 and Y_1 varied between 0.1 and 0.9. For each combination of group size and correlation, we generated 6,400 trials and analyzed them using ANCOVA. To estimate the true power, we calculated the percentage of trials with a statistically significant treatment effect.

3. Results

In the appendix it is shown that for large trials the design factor (variance deflation factor) for ANCOVA is $1-\rho^2$, where ρ is the correlation between Y_0 and Y_1. As a consequence, ANCOVA with $(1-\rho^2)n$ subjects has the same power as *t*-test with *n* subjects. It is straightforward to calculate that for a *t*-test on the change from baseline $(Y_1 - Y_0)$, the design factor is $2-2\rho$: if a *t*-test on Y_1 requires *n* subjects, then a *t*-test on the change from baseline requires $(2-2\rho)n$ subjects.

The design factor can also be used to compare the precision of the analysis methods. If ANCOVA is used instead of *t*-test, this increases the precision by $\sqrt{1-\rho^2}$, that is, the standard error and the length of the confidence interval (CI) of the difference between the treatment groups are reduced by a factor $\sqrt{1-\rho^2}$. When change from baseline is used, the precision changes by a factor $\sqrt{2-2\rho}$. For $\rho > 0.5$ this is a decrease, but for $\rho < 0.5$ it is an increase.

3.1. Accuracy of sample size calculation methods for the t-test

As expected, the sample sizes calculated by Nquery Advisor® were quite accurate. However, the formula $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/(m_B - m_A)^2$ leads to sample sizes with too little power. The thin and bold broken lines in Fig. 1 show the "true" power, when according to the formula it should have been 80% and 90%, respectively. The unbroken lines show the power when the formula $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/(m_B - m_A)^2 + 1$ was used. An increase of one subject per treatment arm, two in total, leads exactly to the required power.

3.2. Accuracy of the sample size calculation method for ANCOVA

The design factor $1-\rho^2$ may not be accurate for small trials. In Fig. 2a, the true power (vertical axis) is plotted against the correlation between Y_0 and Y_1 (horizontal axis).

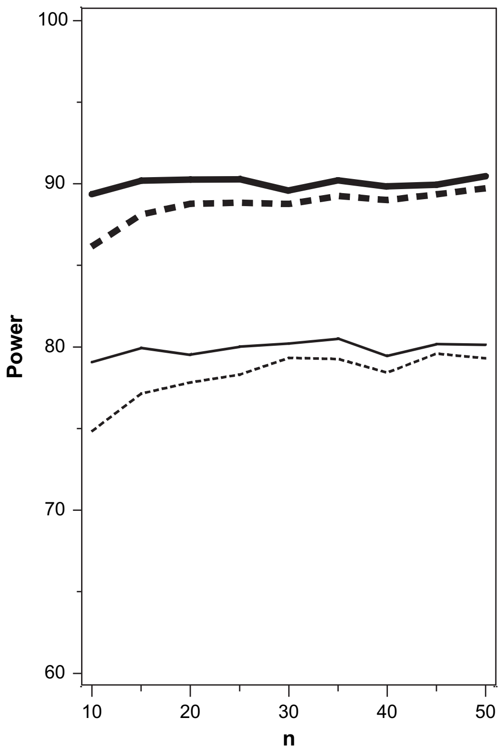

Fig. 1. Accuracy of sample size formula $n = 2(Z_{\alpha/2} + Z_\beta)^2\sigma^2/(m_B - m_A)^2$. The thin and the bold broken lines indicate the true power when the intended (nominal) power was 80% and 90%, respectively. The unbroken lines indicate the true power when one extra subject was added per treatment group.

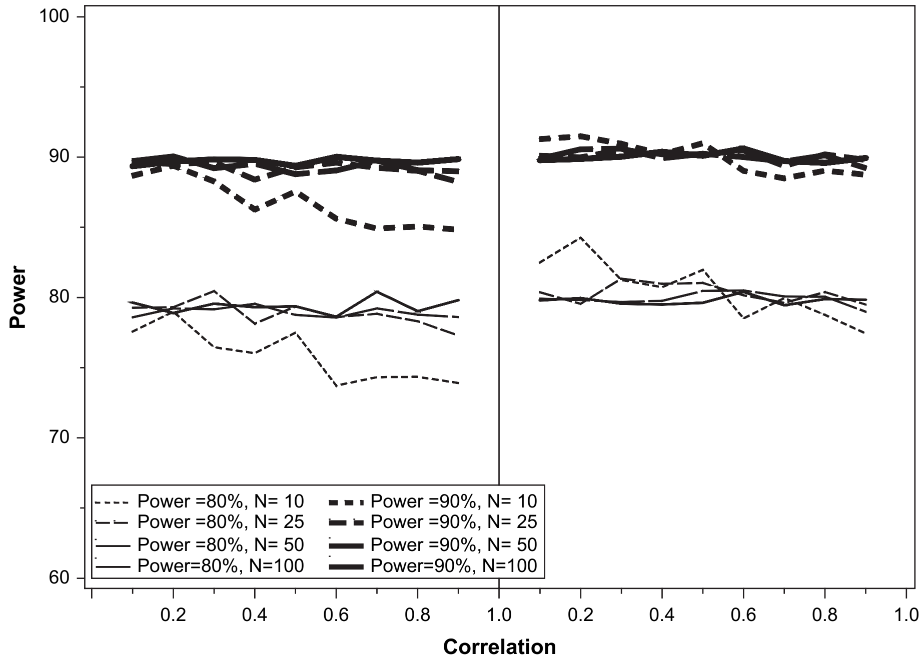

Fig. 2. Accuracy of the sample size calculation method for ANCOVA. (Left) True power when the intended (nominal) power was 80% (thin lines) and 90% (bold lines). (Right) True power after addition of one extra subject per treatment group.

The bold and the thin lines in the figure show the true power when the calculated power based on the design factor and the results of Nquery Advisor® was 90% and 80%, respectively.

It can be seen that for trials with sizes of $n = 25$ per treatment group or higher, the true power was approximately equal to the nominal power. For trials with $n = 10$ per treatment group, the true power was too low. However, adding *one* subject to each treatment group was sufficient to attain the required power, as is illustrated in Fig. 2b.

We repeated the simulation using the formula $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/(m_B - m_A)^2 + 1$ instead of Nquery Advisor® in the first step of the sample size determination method, and found similar results.

The sample size method based on the design factor $1 - \rho^2$ was quite accurate for large sample sizes. By adding one additional subject to each group, it became accurate for all sample sizes.

3.3. Relative efficiency analysis of endpoint, analysis of change from baseline, and ANCOVA

The lines in Fig. 3 show the relative sample sizes that are required when analysis of follow-up scores, analysis of change scores, and ANCOVA are used. The correlation between Y_0 and Y_1 is on the horizontal axis, whereas the vertical axis shows the required sample size. The analysis of follow-up scores is considered the reference method, so the sample size required for this method is set at 100 (thin horizontal line). The thin dashed line shows the sample size requirements when change scores are used. For a low correlation ($\rho < 0.5$), the use of follow-up scores requires a lower sample size than the use of change scores. In contrast, when the correlation is high ($\rho > 0.5$) the use of change scores requires a lower sample size. ANCOVA always has the lowest sample size requirement in comparison

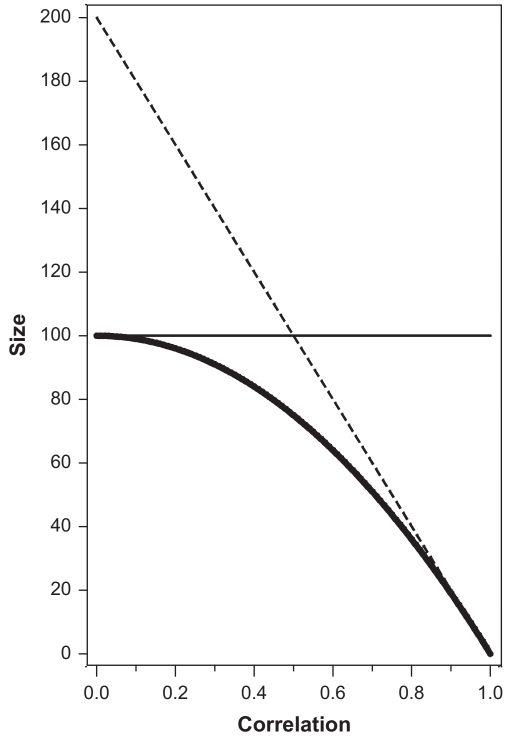

Fig. 3. Relative efficiency of analysis of endpoint, analysis of change from baseline, and ANCOVA. Sample size requirements when endpoint analysis (reference; thin line), change from baseline (dashed line) or analysis of covariance (bold line) is used.

to the other methods (bold line). For correlations between $\rho = 0.2$ and $\rho = 0.8$, the reduction in sample size for ANCOVA versus analysis of follow-up scores ranges between 4% and 64%. For ANCOVA versus analysis of change scores the reduction ranges between 40% and 10%. Above $\rho = 0.8$, the advantage of ANCOVA over the use of change scores is small.

4. Example

As an illustration of the method, we use a randomized placebo-controlled trial, designed to investigate whether treatment with leflunomide and sulfasalazine is more effective than treatment with sulfasalazine alone. Patients with rheumatoid arthritis who had insufficient clinical response to treatment with leflunomide were randomized to leflunomide and sulfasalazine or to placebo and sulfasalazine [5]. The primary outcome measure was the Disease Activity Score (DAS28), which was measured at baseline and after 24 weeks of treatment [6]. It was hypothesized that improvements would occur in both treatment arms, but addition of sulfasalazine would be more effective than a mere switch to sulfasalazine [5].

We calculated the sample size requirements per treatment group using $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/(m_B - m_A)^2 + 1$ assuming a significance level of $\alpha = 0.05$ (two-sided) and a power of $1 - \beta = 0.80$. A difference of 0.6 DAS28 points between the treatments was deemed appropriate, and the SD was estimated to be 1.2 [5,7,8]. The correlation of the DAS28 between subsequent 24-week periods was $r = 0.7$, as determined using the data of 365 rheumatoid arthritis patients with up to 6 years follow-up [9]. In Table 1, the required sample sizes are given for three methods of analysis: t-test on follow-up scores (Y_1), t-test on change from baseline ($Y_1 - Y_0$), and ANCOVA ($Y_1|Y_0$). Sample size requirements for more stringent values of the significance level α and power $1 - \beta$ are shown at the right hand side of the table. With a correlation coefficient between Y_1 and Y_0 of 0.7, the use of ANCOVA leads to a sample size reduction of 13% in comparison to the analysis of the change from baseline and a reduction of 48% versus the analysis of the endpoint scores.

Table 1
Sample sizes for three methods of analysis of the rheumatoid arthritis trial

Error rates	$\alpha = 0.05$, $1 - \beta = 0.80$			$\alpha = 0.01$, $1 - \beta = 0.90$		
Method of analysis	Follow-up scores	Change scores	ANCOVA	Follow-up scores	Change scores	ANCOVA
ρ	N	N	N	N	N	N
0	126	252	126	240	480	240
0.5	126	126	95	240	240	180
0.6	126	101	81	240	192	154
0.7	126	76	64	240	144	122
0.8	126	50	45	240	96	86
0.9	126	25	24	240	48	46

Note: Total sample size ($N = 2n$) for different values of α and $1 - \beta$ (power) and for different values of the correlation ρ between the DAS28 at follow-up and at baseline.

5. Discussion and conclusion

Covariate adjustment increases the power and reduces the sample size in RCTs [1−3,10,11]. Another advantage of covariate adjustment is that it corrects for imbalances that may have occurred despite the randomization [12,13].

We propose a simple method for the sample size calculation when ANCOVA is used: multiply the number of subjects required for the t-test by $1 - \rho^2$ and add one extra subject per group. Then add some additional subjects to compensate for potential missing and non-evaluable observations.

5.1. Other covariates

We have discussed the sample size calculation when the outcome variable measured at baseline is the independent covariate in the ANCOVA. It may be clear that the results in this paper remain valid when another baseline variable is used as covariate.

Our approach can easily be generalized to more than one covariate. The design factor remains the same, but ρ in this case is the multiple correlation between the baseline variables $X_1,\ldots X_k$ and the follow-up outcome Y_1 (within the treatment groups). This correlation may be unknown, but it is at least as large as the largest of the correlations between Y_1 and the individual baseline variables. When data from earlier trials or cohorts are available, the multiple correlation can be estimated by carrying out a linear regression with dependent factor Y_1 and independent factors $X_1,\ldots X_k$. The percentage explained variance R^2 is an estimate of ρ^2. The design factor is $1 - R^2$.

5.2. Selection of covariates

It is important that the covariates that the analysis adjusts for are prespecified in the study protocol because inclusion of variables on the basis of data-dependent selection may lead to spurious results [10,11,13,14]. Data from previous trials on similar treatments and patient populations may be useful for the selection of covariates for the analysis [14]. In general, variables that are thought to be strongly predictive of the outcome are likely candidates, both for inclusion as covariates in the analysis and as stratification or minimization factors in the randomization. If the randomization is stratified for certain variables, these should be included as covariates in the analysis [15].

5.3. Assumptions of the power calculation

The power calculation for a t-test requires assumptions about the SD and the difference between the groups. For

ANCOVA, it also requires an estimate of the correlation between the covariate and the outcome. Similar to the SD, this correlation may be estimated from similar previous trials or from observational data.

When the planned analysis of the trial is ANCOVA, it is assumed that the slopes of the regression lines in both treatment groups are equal (no interaction or treatment modification). However, this assumption may be incorrect, and so may be the estimates of the correlation, SD, or treatment difference. This highlights a disadvantage of an ANCOVA approach versus a *t*-test approach: it depends on more assumptions. If some of these assumptions are incorrect, the chance that the trial is successful may be less than is calculated.

5.4. Conclusion

The use of ANCOVA may considerably reduce the number of patients required for a trial. This should be taken into account when the sample size of a trial is determined. The approach presented in this paper offers a simple method to do so.

Appendix

We assume that in a randomized trial with two treatment groups ($Y_{1,1}$, $Y_{1,2}$,... $Y_{1,n}$, $Y_{2,1}$, $Y_{2,2}$,...$Y_{2,n}$) is the outcome of interest and ($X_{1,1}$, $X_{1,2}$,...$X_{1,n}$, $X_{2,1}$, $X_{2,2}$,...$X_{2,n}$) is the baseline. All pairs ($Y_{\mathrm{i,j}}$, $X_{i,j}$) follow a bivariate normal distribution with known correlation ρ and SDs σ and τ. The mean of $X_{i,j}$ is 0 and the mean of $Y_{i,j}$ is μ_i (i = 1,2). An ANCOVA corresponds to testing equality of the means of $Y_{i,j}$ in the treatment groups, given that ($X_{1,1}$, $X_{1,2}$,...$X_{1,n}$, $X_{2,1}$, $X_{2,2}$,...$X_{2,n}$) = ($x_{1,1}$, $x_{1,2}$,...$x_{1,n}$, $x_{2,1}$, $x_{2,2}$,...$x_{2,n}$). Then the (conditional) distribution of $Y_{i,j}$ has variance $(1-\rho^2)\sigma^2$ and mean μ_i [16].

The power of the *t*-test is based on the variance σ^2, whereas the power of ANCOVA depends on the conditional variance $(1-\rho^2)\sigma^2$. As the latter is a factor $1-\rho^2$ smaller and as the sample size required for a certain power is proportional to the variance, an ANCOVA on $(1-\rho^2)n$ observations has the same power as a *t*-test on n observations.

The reduction of the variance by a factor $1-\rho^2$ also means that the use of ANCOVA increases the precision by $\sqrt{1-\rho^2}$, that is, the standard error and the length of the CI of the difference between the treatment groups are reduced by a factor $\sqrt{1-\rho^2}$.

When ρ and σ are known, the test statistic follows a normal distribution, and it is straightforward to show that the required sample size for a test with type I and II errors α and β is $n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2(1-\rho^2)\sigma^2/(\mu_1-\mu_2)^2$ per group. When ρ and σ are not known, the test statistic follows a *t*-distribution with $2n-2$ degrees of freedom. For large trials the difference between a normal and a *t*-distribution is negligible, but for small trials the formula will underestimate the required number of patients.

References

[1] Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123−4.

[2] Venter A, Maxwell SE, Bolig E. Power in randomized group comparisons: the value of adding a single intermediate time point to a traditional pretest-posttest design. Psychol Methods 2002;7:194−209.

[3] Van Breukelen GJ. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. J Clin Epidemiol 2006;59:920−5.

[4] Armitage P, Berry G. Statistical methods in medical research. 3rd edition. Oxford: Blackwell Science; 1994.

[5] Dougados M, Emery P, Lemmel EM, Zerbini CA, Brin S, Van Riel P. When a DMARD fails, should patients switch to sulfasalazine or add sulfasalazine to continuing leflunomide? Ann Rheum Dis 2005;64: 44−51.

[6] Prevoo MLL, Van 't Hof MA, Kuper HH, Van Leeuwen MA, Vande Putte LB, Van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44−8.

[7] Van Gestel AM, Haagsma CJ, Van Riel PLCM. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. Arthritis Rheum 1998;41:1845−50.

[8] VandePutte LB, Atkins C, Malaise M, Sany J, Russell AS, Van Riel PL, et al. Efficacy and safety of adalimumab as monotherapy in patients with rheumatoid arthritis for whom previous disease modifying antirheumatic drug treatment has failed. Ann Rheum Dis 2004;63:508−16.

[9] Welsing PM, Van Riel PL. The Nijmegen inception cohort of early rheumatoid arthritis. J Rheumatol Suppl 2004;69:14−21.

[10] Hernandez AV, Steyerberg EW, Habbema DF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. J Clin Epidemiol 2004;57:454−60.

[11] Hernandez AV, Eijkemans MJC, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann Epidemiol 2006;16: 41−8.

[12] Twisk J, Proper K. Evaluation of the results of a randomized controlled trial: how to define changes between baseline and follow-up. J Clin Epidemiol 2004;57:223−8.

[13] European Agency for the Evaluation of Medicinal products (EMEA), Committee for Proprietary Medicinal Products (CPMP). Points to consider on adjustment for baseline covariates. London: EMEA; 2001.

[14] Raab GM, Day S, Sales J. How to select covariates to include in the analysis of a clinical trial. Control Clin Trials 2000;21:330−42.

[15] Bland JM. Sample size in guidelines trials. Fam Pract 2000;17(Suppl 1):S17−20.

[16] Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. London: McGraw-Hill; 1974. 167 p.