Published in final edited form as:
J Biopharm Stat. 2015 ; 25(4): 812–829. doi:10.1080/10543406.2014.920858.

An Evaluation of Statistical Methods for Analyzing Follow-Up Gaussian Laboratory Data with a Lower Quantification Limit

John M. Karon[a], **Ryan E. Wiegand**[b], **Janneke H. van de Wijgert**[c], **Peter H. Kilmarx**[b]

[a]Apex Systems, Inc., Richmond, Virginia, USA

[b]Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[c]Institute of Infection and Global Health, University of Liverpool, United Kingdom

Abstract

Laboratory data with a lower quantification limit (censored data) are sometimes analyzed by replacing non-quantifiable values with a single value equal to or less than the quantification limit, yielding possibly biased point estimates and variance estimates that are too small. Motivated by a three-period, three-treatment crossover study of a candidate vaginal microbicide in HIV-infected women, we consider four analysis methods for censored Gaussian data with a single follow-up measurement: nonparametric methods, mixed models, mixture models, and dichotomous measures of a treatment effect. We apply these methods to the crossover study data and use simulation to evaluate the statistical properties of these methods in analyzing the treatment effect in a two-treatment parallel-arm or crossover study with censored Gaussian data. Our simulated data and our mixed and mixture models consider treated follow-up data with both the same variance as the baseline data and an inflated variance. Mixed models have correct type I error, the best power, the least biased Gaussian parameter treatment effect estimates, and appropriate confidence interval coverage for these estimates. A crossover study analysis with a period effect can greatly increase the required study sample size. For both designs and both variance assumptions, published sample size estimation methods do not yield a good estimate of the sample size to obtain stated power.

Keywords

mixed models; mixture models; laboratory data; quantification limit

1. INTRODUCTION

Sexual transmission is the principal mode of HIV transmission in much of the world. Transmission risk increases with increasing HIV viral load in reproductive tract fluid (cervico-vaginal or seminal fluid) (Baeten et al., 2011). A few studies (Dunne et al., 2008; McLean et al., 2010) have evaluated therapeutic interventions to reduce cervico-vaginal and seminal fluid viral loads.

Correspondence to: John M. Karon, 2505 Elfego Road NW, Albuquerque, New Mexico 87107-3010, JKaron@earthlink.net, 505.342.5639; fax 509.355.6747.

HIV assays have become more sensitive, but a major problem in data analysis remains the substantial proportion of genital tract samples with HIV viral loads below the quantification limit (i.e., data are truncated, commonly called left-censored in the literature). For example, in a study of African serodiscordant couples, HIV-1 RNA was below the detection limit for 40% of endocervical swab samples and 43% of semen samples (detection limits, 240 copies per swab and per mL, respectively: Baeten et al., 2011). In a three-period, three-treatment crossover trial, balanced on period and treatment ("treatments" were no product, placebo gel, and Carraguard gel), of a vaginal microbicide (Carraguard) in 60 HIV-infected women, 34% – 51% had unquantifiable cervico-vaginal lavage (CVL) HIV-1 RNA (< 80 copies/mL) at the start of the three treatment periods (McLean et al., 2010).

The primary interest in these studies and future studies is the reduction of reproductive tract HIV RNA associated with treatment. Laboratory data including values below the quantification limit were initially analyzed by simply replacing the censored values by Q (the quantification limit) or Q/2 and using an analysis appropriate for continuous data. This practice has continued (Baeten et al., 2011) despite the development of more appropriate analysis methods and the publication of simulations demonstrating bias in the point estimate and underestimation of variability. For example, Hughes (1999) generated longitudinal data from a random effects model and found negative biases in the estimate of a linear trend and in the estimates of the variances when non-quantifiable values are replaced by Q or Q/2. Lubin et al. (2004) generated univariate data from a Gaussian distribution and found a negative bias in the estimated mean and a confidence interval with less than the nominal coverage when the non-quantifiable values are replaced by Q/2 unless only 10% of the data were censored.

The literature contains many statistical methods for analyzing censored laboratory data. We consider nonparametric methods, mixed models, mixture models, and models for a change from baseline to follow-up. There are reviews comparing analysis procedures (Journot et al., 2001; Senn et al., 2012; Lachenbruch, 2001) and point estimate and variability bias (Nie et al., 2009), but there is no comprehensive comparison of the power of these methods including the models summarized above.

After describing the statistical methods we evaluated, we analyze data from the Carraguard trial to demonstrate potential differences among the conclusions obtained from these statistical methods. We use simulation to evaluate the size and power of tests using these methods for a two-treatment placebo-controlled study using parallel and two-period crossover designs. We evaluate both the validity of published sample size estimation procedures to achieve desired power and the bias and confidence interval coverage of the Gaussian parameter estimate for the treatment effect in mixed and mixture models. In the Discussion, we comment on the choice between parallel arm and crossover study designs. Additional details are in a supplementary report.

2. STATISTICAL METHODS

We assume that, in each treatment period (crossover studies have multiple treatment periods), Gaussian laboratory data are obtained at baseline and at a single follow-up time.

Let y_{it} be the assay value for the ith person at the tth time during a parallel design study with a single follow-up time ($t = 0$ is baseline; t = 1 is follow-up). We use the same notation in each period for a cross-over study with a single follow-up time in each period. Let Q be the lower limit of the quantifiable values after any transformation needed to obtain Gaussian data. Assay results less than Q are censored; a difference between or among groups is a treatment effect. We may distinguish between censored assay results that are non-detectable (ND) and results that are detectable but non-quantifiable (NQ). For HIV viral loads, an assay result that is NQ means that the assay provided a positive numerical result that is less than the quantification limit Q and hence is too imprecise for the value to be used in analyses (virus is present at a low level, but the actual level is uncertain). An assay result that is ND means that no virus could be detected.

2.1 NONPARAMETRIC PROCEDURES

Nonparametric procedures can be used when analyzing only the follow-up values. For the parallel design, we consider the Wilcoxon two-sample test and, with ND and NQ values combined, Gehan's modification of this test for data that are right censored at a single fixed value (Gehan, 1965). Let the observed values be denoted by y_i. To compute p-values for the Gehan test, we set all values below the quantification limit (both ND and NQ values) to a value less than this limit, defined a new value z_i as $2 – y_i/\max(y_i)$ (so that the values below the quantification limit are equal and are the largest values), and used the SAS LIFETEST procedure to analyze the z_i (defining the NQ and ND values as censored). For the crossover design, we consider the Friedman test (analysis of variance based on ranks: Randles and Wolfe, 1979), with persons representing blocks. Note that none of these analyses estimate the magnitude of a treatment effect.

2.2 MIXED MODELS AND MIXTURE MODELS: MAXIMUM LIKELIHOOD

Procedures have been proposed that yield likelihoods, so that maximum likelihood estimators can be obtained. Hughes (1999) developed an algorithm for imputing single values from an EM algorithm. Lyles et al. (2000) modified the numerical procedure in Hughes' random effects model by integrating out the random effects to obtain a likelihood function that can be maximized numerically. Thiébaut and Jacmin-Gadda (2004, 2007) provide code using the SAS® NLMIXED procedure to maximize such a likelihood.

To define a mixed model, assume that we have n values, indexed so that the first m values are censored at Q (NQ or ND). Let y_i be the value of the ith observation, and let $\boldsymbol{x}_i$ be the covariates for this observation; $\boldsymbol{x_i}$ includes an intercept term. For a design evaluating a single treatment, $\boldsymbol{x_i}$ contains an indicator for a value obtained at follow-up, an indicator for active treatment, and, for the crossover design, and indicator for period 2. Let φ and Φ be the standard Gaussian density and cumulative distribution functions, respectively. Let the estimated mean for the ith observation be $\mu_i = \boldsymbol{x}'_i\boldsymbol{\beta}$, where the vector $\boldsymbol{\beta}$ contains the coefficients to be estimated, and let σ be the standard deviation of the distribution of the y_i. The likelihood to be maximized is

$$L_{mixed} = \prod_{i=1}^{m} \Phi\left(\frac{Q - \mu_i}{\sigma}\right) \prod_{i=m+1}^{n} \frac{1}{\sigma}\phi\left(\frac{y_i - \mu_i}{\sigma}\right) \quad (1)$$

Lynn (2001) proposed maximizing this likelihood. We define the covariates $\boldsymbol{x_i}$ and parameters $\boldsymbol{\beta}$ in Section 2.4.

Moulton et al. (2002) define a mixture likelihood with a logistic model for the probability p_i that the assay result is quantifiable. Let f and F be the probability distribution and cumulative distribution functions, respectively, of the laboratory values, and let m and n be defined as above. For observations with covariates $\boldsymbol{x}_i$, let $y_i = \boldsymbol{x}'_i\boldsymbol{\beta} + \sigma\varepsilon_i$ where the ε_i have a standard normal distribution. Let $w_i = (y_i - \boldsymbol{x}'_{\boldsymbol{i}}\boldsymbol{\beta})/\sigma$ and let w_i^* be the value of w_i when y_i = Q. The likelihood is

$$L_1 = \prod_{i=1}^{m} (1 - p_i + p_i F(w_i^*)) \prod_{i=m+1}^{n} [p_i f(w_i)/\sigma] \quad (2)$$

The first product represents observations that are censored. The term $p_i F(w_i^*)$ represents the probability that a person whose true laboratory value is quantifiable has an assay value that is censored; this likelihood corresponds to a model for survival data, with long-term survivors who are censored at the end of a study (Farewell, 1983). The second product represents observations that are quantifiable. We chose f and F to be the density and cumulative distribution function, respectively, for the Gaussian distribution. Because our simulations showed that this likelihood gives a type I error that is much too large (Supplementary Report, Tables 6 and 9), as did the same likelihood deleting the long-term survivors term $p_i F(w_i^*)$, we do not consider these likelihoods. We found that the likelihood for uncensored data

$$L_2 = \prod_{i=1}^{m} (1 - p_i) \prod_{i=m+1}^{n} p_i f(w_i) \quad (3)$$

gives appropriate Type I error rates in some cases, as reported in the summary of simulations results in Section 4. In our analyses and simulations, we do not assume that the corresponding parameters in the logistic and Gaussian means of (3) are equal; we define these means for the choices of models and data used in Section 2.4. We choose the parameters so that negative values for both the Gaussian and logistic treatment coefficients indicate a favorable treatment effect. For longitudinal data, we follow Moulton et al. by using the robust covariance matrix estimated by

$$V\left[\sum_{i=1}^{n} S_i S_i'\right]V$$

where V is the inverse Hessian and $\boldsymbol{s}_i$ is the contribution of the i^th^ observation to the score vector from the logarithm of the full likelihood function.

2.3 LOGISTIC REGRESSION: MODELING A CHANGE DEFINED BY A MAGNITUDE OF INTEREST

A binary definition of a treatment effect would simplify the data analysis and might directly address whether a desired treatment effect was achieved. In the Carraguard study, clinicians believed that a clinically significant change in viral load was a decrease from baseline to follow-up of 0.5 on the $\log_{10}$ scale if both results are quantifiable, a decrease from quantifiable to NQ or ND, or (if we consider NQ as different from ND) a decrease from NQ to ND. A meaningful increase can be defined analogously.

We evaluated whether such a definition would have comparable power to other analyses. We use logistic regression to analyze this binary definition of a meaningful treatment effect. We model the logit of no decrease, so that a negative estimate corresponds to a desired treatment effect, as in the mixed models and mixture models. For longitudinal data, we use generalized estimating equation (GEE) logistic regression models. For the AB:BA crossover design with a dichotomous treatment effect definition, Schouten and Kester (2010) propose an analysis similar to a McNemar matched pairs test. We implemented their proposal with a similar definition of a treatment effect but assuming that any decrease in viral load from baseline to follow-up represented a treatment effect.

2.4 METHODS CONSIDERED

We consider all of the nonparametric methods described above. We implemented two mixed models using the likelihood L_{mixed} as proposed by Thiébaut and Jacqmin-Gadda (2004, 2007), one assuming that the variance of assay values is unaffected by treatment, and a second model allowing this variance to be affected by treatment (as suggested by the Carraguard data). The mixed models fit to Carraguard data and simulated crossover data consider using follow-up values only and both baseline and follow-up data; those fit to simulated parallel design data use both the baseline and follow-up values. For the crossover design, we also consider mixed models with period effects. We evaluated mixture models with both variance assumptions; we consider models using follow-up data only as well as both baseline and follow-up data. For the crossover design simulation analyses, we implemented the logistic regression models for a change defined by a magnitude of interest as GEE models and also used the Schouten-Kester procedure.

In our models, we coded treatments and periods as indicator variables. In simulations, the placebo is the reference treatment; in analyzing the Carraguard data, no-treatment is the reference. We assume that the mean value at baseline in a parallel design study is the same in all treatment groups, with the corresponding assumption at the start of each period in a cross-over study. We also assumed that the time elapsed from baseline to follow-up has the same effect on the mean in each treatment group (any treatment effect is in addition to this time effect). For both the Carraguard data and the simulations of a crossover design, we assumed that there were no carryover effects. For cross-over trial analyses with period effects, period 1 is the reference.

These assumptions yield the following models for the means μ in our simulations of a treatment versus placebo study; analyses of the three-period, three treatment Carraguard study use the corresponding extensions. In the mixture models, the Gaussian and the logistic means have the same form; the corresponding coefficients are not assumed to be equal. In the following models, x is an indicator variable for treatment, δ_1 is an indicator variable which is 1 for a value obtained at follow-up (t = 1), β_x estimates the treatment effect, and β_1 estimates the change from baseline to follow-up in the reference group.

For the parallel design, the mixed models use both the baseline and the follow-up results. The model for the mean for the ith observation in the likelihood (1) is

$$\mu_i = \alpha + \delta_{1i}\,\beta_x x_i + \beta_1\delta_{1i} + a_i \tag{4}$$

For this design analyzed with a mixture model using only the follow-up results

$$\mu_i = \alpha + \ \beta_x x_i \tag{5}$$

For the crossover design, we assumed that the elapsed time from baseline to follow-up has the same effect on the mean in all periods. Let the indicator variable T be 1 in the period in which treatment was used. The mixed models for the crossover design using both baseline and follow-up values without period effects use

$$\mu_i = \alpha + T\delta_{1i}\,\beta_x x_i + \beta_1\delta_{1i} + a_i \tag{6}$$

The mixed models using only the follow-up data use

$$\mu_i = \alpha + T\,\beta_x x_i + a_i \tag{7}$$

In (4), (6), and (7), a_i is a random effect with a Gaussian distribution. For both designs and both data choices, the mixture models use the same mean without the random effect. The GEE logistic model for a significant decrease uses the model (5). For the analysis of a crossover study with a period effect, we added an indictor variable for period 2 to the means in (6) and (7).

To clarify the definitions of these means, for a follow-up design and a cross-over design (without period effects) using both baseline and follow-up values, the mean value at baseline (and at the start of period 2 for the cross-over study) is α; the mean value at follow-up unaffected by treatment is $\alpha + \beta_1$; the mean at follow-up affected by treatment is $\alpha + \beta_1 + \beta_x$. For the cross-over study analyzed using a mixture model and only the follow-up data, the mean values unaffected and affected by treatment are α and $\alpha + \beta_x$, respectively.

In our simulations, we evaluated the significance of the treatment effect β_x from the Wald chi square for the mixed models and the logistic regression models. We used the likelihood ratio test (compared to a model without treatment) for the mixture models; these tests have 2 and 3 degrees of freedom for models assuming variances unaffected and affected by treatment, respectively (because the treatment effect coefficients for the Gaussian and logistic portions of the model need not be equal). For the mixed and mixture models, we required that the Hessian matrix have full rank (for the mixture models, the Hessian for both the full and reduced models). For the mixture models, we also required that the likelihood ratio statistic be positive and that the magnitude of the standard error for the logistic coefficient be at most 5 (to discard results with unstable estimates of the logistic portion of the model, resulting when a data set had few non-quantifiable observations). A significant model favored treatment or placebo use only when the coefficient(s) had the appropriate sign(s) (for the mixture models, both the Gaussian and logistic coefficients); for nonparametric models, the direction of the effect was based on the medians in the treatment and placebo groups.

In the Carraguard study, we assessed the variation among treatments using mixed models and mixture models with the likelihood ratio chi-square statistic with 4 and 5 degrees of freedom for models assuming variances unaffected and affected by treatment, respectively. We used the Wald chi-square statistic with 2 degrees of freedom (computed using SAS PROC IML) to assess the significance of variation among treatments for the GEE models.

We used SAS version 9.1 for our analyses of the Carraguard data and versions 9.2 and 9.3 for the simulations. We used the following procedures: NPAR1WAY for the Wilcoxon test, FREQ for the Friedman test, LIFETEST for Gehan's test, NLMIXED for mixed models, NLMIXED to maximize the likelihood for mixture models, and LOGISTIC and GENMOD for logistic regression for parallel design and crossover design data (GEE models), respectively.

3. ANALYSIS OF THE CARRAGUARD PHASE I STUDY

The Centers for Disease Control and Prevention (CDC) and the Population Council conducted a phase I study of a potential vaginal microbicide, Carraguard, in HIV-infected women in Chiang Rai, Thailand, during March 2003 to June 2004 as a three-treatment, three-period crossover trial (McLean et al., 2010). The "treatments" were no product, a placebo methylcellulose gel, and a Carrageenan-based Carraguard gel (both administered double-blinded). Each assessment period lasted 28 days. Sixty women were enrolled, with 10 randomly assigned to each of the six possible treatment sequences.

For all three treatment arms, a cervico-vaginal lavage (CVL) sample was obtained on days zero (before product use), seven (after the final daily use of a product), and 14 of each menstrual cycle. Thus there was a washout period of approximately 21 days between treatment periods. CVL samples were analyzed for HIV-1 RNA using standard procedures (McLean et al., 2010). Data were reported as copies per mL of CVL fluid; all numerical values in this section use these units. The lower limit for quantification was 80 copies/mL; samples with no evidence of HIV-1 RNA were classified as non-detectable.

A first step in the analysis is a test for variation of the HIV-1 RNA values at day seven, or the change from day zero to day seven, among treatments. We show the results of analyses using the relevant statistical methods defined in the previous section. The mixed models and mixture models combine the NQ and ND values; the nonparametric tests and the logistic regression models consider combining these values as well as separating the ND values. For each woman, we restricted data to the 163 woman-cycles with an HIV-1 RNA result at both days zero and seven without antiretroviral treatment in that cycle. Six woman contributed data at only one cycle (5 started antiretroviral treatment in cycle 2); 5 contributed data at two cycles (one started treatment in cycle 3). Only one woman was lost to follow-up (McLean et al., 2010).

Table 1 shows the distributions of the $\log_{10}$ HIV-1 RNA values in copies per mL on days zero, seven, and 14 of each cycle (the published data are total copies, hence different from Table 1). The HIV-1 values were quantifiable for 42% to 66% of the samples on individual days. The descriptive statistics for day zero suggest that the HIV-1 RNA values were lower at the start of cycles two and three than at the start of the study. The Carraguard and placebo gel treatment effects on day seven are similar. Carageenan is not absorbed in the vagina and the half-life of its activity has not been estimated, but comparing the day 14 and day zero results suggests that even a 7-day washout period was adequate. Q-Q plots and formal statistical tests show that the quantifiable $\log_{10}$ CVL values have a normal distribution (Supplementary Report).

Results from analyzing these data (excluding period effects) are summarized in Table 2. The models use the means in (4)–(7), as appropriate, but with two parameters for treatment (placebo gel and Carraguard). The Friedman tests and the mixed models using only the data from day seven provide strong evidence for variation among the treatments in their effects on HIV-1 CVL RNA. The mixed models using data from both days zero and seven provide weak evidence for such variation. The mixture models and the GEE logistic regression models for a meaningful decrease do not suggest such variation. Because only 17 of 163 changes from day zero to day seven were classified as a meaningful increase in HIV-1 CVL RNA (including changes from ND to NQ), we analyzed only indicators for a meaningful decrease in the GEE models. The coefficient estimates for the placebo gel and Carraguard treatment effects obtained from the mixed models using data only from day seven suggest that both the placebo gel and Carraguard reduce HIV-1 CVL RNA by approximately 0.5 on the $\log_{10}$ scale. For comparison, for the 45 woman-cycles using placebo gel or Carraguard during which both the day zero and day seven values were quantifiable, the median decrease in $\log_{10}$ RNA HIV-1 during this interval was 0.42 copies/mL. The corresponding coefficient estimates from the mixture models are smaller in magnitude. We show why the GEE models have poor power to detect a treatment effect at the end of the presentation of simulation results for the parallel arm design.

We used heuristic methods to evaluate whether the mixed models with equal variances fit the data. We obtained the predicted value for each woman at each time from the PREDICT statement in the SAS NLMIXED procedure. For the models using day seven only and both days, 15 of 163 and 41 of 326 observations, respectively, had predicted values that disagreed with the observed values with respect to quantification. For the model using day seven only, agreement between observed and predicted numbers was good except for the periods using no product, which had seven of the 15 misclassifications (all misclassified as NQ: Supplementary Report, Table 3). In both models, the Gaussian q-q plot of the prediction errors for the observed quantifiable values supported Gaussian errors (Supplementary Report, Figure 1), as did the p-values from three tests (the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling tests) (Supplementary Report, Table 2). Based on these results, we prefer to fit a mixed model to day seven data only and may accept the conclusions from that model.

Mixed models containing period effects had likelihood ratio chi-square statistics, treatment coefficient estimates, and standard errors for these estimates that were very similar to the corresponding values in Table 2 (Supplementary Report, Table 4).

4. SIMULATION EVALUATIONS OF POWER FROM ALTERNATIVE ANALYSES

We used 4000 simulations of two-treatment designs with a Type I error of 0.05 for each analysis method to evaluate whether test size, the power of the tests with appropriate size to detect a treatment effect, and (for the mixed models and mixture models) the bias of the Gaussian parameter estimate of the treatment effect and the coverage of a 95% confidence interval for this estimate. Half-lengths of the confidence intervals for percent power are 1.4 and 1.6 for 80% and 50% power, respectively.

Our simulations used values suggested by the three-period, three-treatment Carraguard study (see the Supplementary Report). For each person, we used the SAS macro MVN in PROC IML to generate multivariate Gaussian data (bivariate for the parallel design, quadrivariate for the crossover design) with standard deviation 1 in the absence of treatment and an exchangeable covariance matrix with correlation 0.6. We assumed that the mean treatment effect was a reduction of 0.5 and that the quantification limit was 1.9 ($\log_{10}$ 80). Based on the Carraguard data, the probabilities that an NQ value was ND were 0.70 and 0.55 for treated and untreated values, respectively. We considered untreated mean values of 2.0, 2.5, 3.0, and 4.0; for a mean of 2.5, we evaluated the size of each test in the absence of treatment. We evaluated the performance of the analysis methods for standard deviations (σ_{Trt}) of 1.0 and 1.4 for the treated values at follow-up. For the models, power calculations were restricted to the models that converged. For a baseline mean of 2.5, the probability of an NQ result at baseline is 0.28; at follow-up in a treated group, the corresponding probabilities are 0.46 and 0.47 for σ_{Trt} equal to 1 and 1.4, respectively. See Supplementary Report, Table 5, for these probabilities for other baseline means.

4.1. PARALLEL DESIGN

The nonparametric analyses (the Wilcoxon test) used only the follow-up data. The mixed models and the logistic models for a meaningful change used both baseline and follow-up data. For both assumptions on the variance of the treated values, we used the likelihood L_2 to fit mixture models using follow-up data only as well as using both baseline and follow-up data. We used the mean (4) for the mixed models and (4) and (5) withut the random effects term for the mixture models.

We estimated sample sizes using Lachenbruch's method for a "two-part" estimator in which the censored and quantifiable observations are analyzed separately (Lachenbruch, 2001). This procedure estimates the non-centrality parameter for the two degree-of-freedom chi-square test statistic when there is a common variance in the two treatment arms; it can be generalized to different variances in the two arms. For a treatment effect of 0.5, a baseline mean of 2.5, and a quantification limit of $\log_{10}$ 80, the estimated sample sizes to obtain 80% power are 64 and 73 persons per arm with σ_{Trt} equal to 1.0 and 1.4, respectively. A two-sample t-test gives the same estimate for the equal variance case.

With no treatment effect, all tests and models except the mixture models had Type I error rates of approximately 5% for both values of σ_{Trt} with 64 persons per arm and a baseline mean of 2.5. Except for the model using baseline data with $\sigma_{Trt} = 1$, the mixture models had Type I error rates of 15% to 57% (Supplementary Report, Table 6).

Table 3 shows power results for a treatment effect of 0.5. In general, power decreased as the baseline mean decreased. With NQ and ND values combined, the Wilcoxon and Gehan tests had nearly identical power. When $\sigma_{Trt} =1$, the mixed models achieved greater than 80% power with 64 persons per arm, even for a baseline mean of 2.0. When $\sigma_{Trt} =1.4$, we found that 90 persons per arm are required to achieve approximately 80% power for a mixed model allowing unequal variances when the baseline mean is 2.5; the mixture models allowing unequal variances had poor power with this baseline mean, possibly explained by the fact that on 34% to 60% of the simulations, the likelihood ratio test is significant but the treatment coefficients had opposite signs. The GEE logistic regression models had lower power than the mixed models, as explained below. We did not investigate why some mixture models did not converge, but we expect it was a result of few non-quantifiable values or quantifiable values (when the baseline mean is large or small, respectively). We will see that the same phenomenon occured (but more frequently) with crossover study data.

With $\sigma_{Trt} = 1$ and 64 per arm, the median values of the treatment coefficient estimates and their standard errors from the mixed models were approximately −0.50 and 0.17, respectively, for all baseline means (the latter, slightly larger for a baseline mean of 2.0). With $\sigma_{Trt} = 1.4$ and 90 per arm, the median treatment effect coefficients from the mixed models with unequal variance parameters were −0.40 to −0.50, decreasing in magnitude as the baseline mean decreased. In both cases, the medians of the mixture model Gaussian coefficients had positive bias (substantial bias for baseline means of at most 3: Supplementary Report, Table 7). The treatment effect estimates from most of the mixed models with unequal variances had appropriate coverage for the treatment effect (the exception is a baseline mean of 2.0, with coverage of approximately 92%). Most of the mixture model estimates had poor coverage (Supplementary Report, Table 8).

For Gaussian data, we can compute the probability of a meaningful decrease (as defined above, with NQ and ND values combined), given the true treatment effect and decrease required if both pre-and post-treatment values are quantifiable (Supplementary Report). The results are summarized in Table 4 using the parameter values from the simulation study. This table also shows the power to detect a difference between decreases in the study groups with n=64 in each group as well as the estimated number in each group required to obtain a test with 80% power. For baseline means of 2.5 and 2.0, plausible values for $\log_{10}$ CVL HIV-1 RNA, treatment effects of approximately 0.75 and >1.0, respectively, are needed to have 80 percent power to detect a difference when the treatment effect is 0.5, a difference of 0.5 between quantifiable values is meaningful, and there are 64 in each group. We also did these computations for correlations $\rho = 0.5$ and 0.7; the results were very similar.

4.2. CROSSOVER DESIGN

We evaluated use of follow-up data only using two Friedman tests (ND combined with and separate from NQ), mixed models, and mixture models. We evaluated the use of both baseline and follow-up values using the Schouten-Kester test, mixed models, a mixture model, and two GEE logistic regression models for a meaningful decrease (ND combined with and separate from NQ). We evaluated both the mixed models and the mixture models with σ_{Trt} equal to 1.0 and 1.4. We used the means (4), (5), (6), and (7) (without random effects terms) as appropriate, for these models.

A potential problem with the crossover design is the presence of period and carryover effects. We also generated data with a period effect and evaluated power when there is a period term in the model; we did not consider carryover effects. Senn (2002) advocates ensuring that the washout period is long enough that carryover effects are not necessary and gives reasons for never including a carryover effect in the model.

Senn (2002, Section 9.5.2) provides SAS® code for estimating the sample size required in an AB:BA crossover study with Gaussian data that are quantifiable. For our simulated data, the required total sample size estimates for 80% power are 28 (14 per treatment sequence) and 44 (22 per sequence) for σ_{Trt} equal to 1.0 and 1.4, respectively. Since the data are assumed to be quantifiable, the sample size estimates depend on the relation between the treatment effect and the standard deviation but not on the mean of the baseline distribution. Chu et al. (2006) derived the power for testing the hypothesis that the means of the underlying distributions are equal in a two-treatment mixture model for log-normal data. Their expressions (assuming equal or unequal variances) use the expected proportions of censored values in the two study arms to compute the total sample size from the sample size derived from the means in the truncated distributions of the quantifiable values and the desired power. With a baseline mean of 2.5, a treatment effect of 0.5, and a standard deviation of 1 (unaffected by treatment), their procedure estimates a sample size of 58 per arm. Since this is based only on comparing quantifiable values, this estimate is likely to be too large.

With no treatment effect, the nonparametric tests, mixed models (except for the model using follow-up data only and assuming equal variances when σ_{Trt}=1.4), and logistic regression models had Type I error rates of approximately 5% for σ_{Trt} equal to both 1.0 and 1.4 with 16 persons per arm and a baseline mean of 2.5. Among the mixture models, the Type I error rate was approximately 5% only for the model using only follow-up data and assuming equal variances with σ_{Trt}=1 (Supplementary Report, Table 9).

Table 5 shows power results for a treatment effect of 0.5 for simulations with σ_{Trt} equal to 1.0 and 1.4 (we chose sample sizes to obtain 80% power with an appropriate mixed model when the baseline mean is 2.5). In general, power decreased as the baseline mean decreased. The Friedman tests had poorer power than the appropriate mixed model. The mixed models had the best power (using equal and unequal variance parameters, when the simulated variances are equal and unequal, respectively); the models using follow-up data only had much better power than the models using both baseline and follow-up data. The mixture model and analyses based on a dichotomous definition of improvement (the GEE logistic regression models and the Schouten-Kester analyses) had poor power.

The median treatment coefficient estimates from the mixed models were approximately −0.50 for all baseline means for $\sigma_{Trt} = 1$, whether or not the variances of treated values were assumed to be equal. The losses of power referred to above result from increases in the standard errors of the estimates of the treatment effect as the baseline mean decreases. For $\sigma_{Trt} = 1.4$ the mixed models allowing for unequal variances gave nearly unbiased estimates of the treatment effect from using the follow-up data only but positive biased estimates (magnitudes that are too small) and standard errors which increased as the baseline mean decreased if the baseline data were used. Most of the median mixture model Gaussian treatment coefficient estimates had large positive bias. See Table 10 in the Supplementary Report. As with the parallel design simulations, the mixed models with unequal variances had appropriate coverage (93.9% to 97.0%) for the treatment effect except for one model (Supplementary Report, Table 11).

We also simulated data with a baseline mean of 2.5, a treatment effect of 0.5, and period effects of 0, −0.3, and −0.3 (with standard deviations of the period effects of 0, 0.3, and 0.1, respectively). The power estimates from analyzing 4000 simulations of these data using the nonparametric procedures and the mixed models are summarized in Table 6; all mixed models include period effects (including those for data with a period effect of 0). We also evaluated the Schouten/Kester test, mixture models, and GEE logistic models; as in Table 5, these had poor power. When there are period effects and σ_{Trt} equals 1.0, it required 35 persons per arm to obtain approximately 80% power from mixed models using equal variances (compared to 16 per arm when there are no period effects in the data, as shown in Table 5); models using baseline data have slightly better power than those using only follow-up data. When there are period effects and σ_{Trt} equals 1.4, the mixed models using unequal variances had only 57% to 66% power with 50 persons per arm (compared to power of approximately 80% with 30 per arm using follow-up data only when there are no period effects in the data, as shown in Table 5); again, models using baseline data had slightly better power than those using only follow-up data. This loss of power in the presence of a period effect that reduced the baseline mean in the second period was the result of treatment effect estimates with positive bias (smaller magnitude), except for models using the baseline value when σ_{trt}=1.0, and some increase in the standard error of these estimates (Supplementary Report, Table 12).

5. DISCUSSION

We evaluated selected statistical methods for analyzing study designs with two “treatments” in which the endpoint is the reduction in the quantity of interest, assessed by an assay yielding Gaussian data with a lower quantification limit, at a single follow-up time. Our results are relevant for the analysis of HIV viral load data (we know of no other clinical assays with quantification limits). Most current HIV treatment clinical trials, e.g. Molina et al. (2008), define the endpoint to be a non-detectable viral load; our results do not apply to such trials.

Our simulations and analyses of the Carraguard data found that mixture models and analyses based on dichotomous classification have much poorer power than mixed models and nonparametric tests. A more thorough comparison of analyzing cross-sectional data reached the same conclusion about mixed models versus mixture models (Wiegand et al., submitted). Our results concerning analyses based on a dichotomous outcome concur with prior research giving substantive reasons for not using such a measure of efficacy (Federov et al., 2009; Senn, 2003). The mixture models have other disadvantages, including excessive Type I error for some models and the need to write code to compute the sandwich estimator of variance to obtain standard errors of estimates of the treatment effect when analyzing longitudinal data. When the mixture model likelihood ratio test for a treatment effect is significant, it is relatively common for the Gaussian and logistic parameter estimates to disagree with respect to the direction of the treatment effect. As might be expected, our simulations also show that power may decrease as the proportion of non-quantifiable observations increases. The current lower quantification limit for HIV viral load in the CDC-Atlanta laboratory is 50 copies (Clyde Hart, personal communication, September 2013); the quantification limit is laboratory dependent. Specific power results will depend on the quantification limit.

In our simulations, mixed models (allowing for unequal variances when the data have this property and using follow-up data only for the crossover design) had better power than the nonparametric tests as well as the advantage of yielding an estimate of the treatment effect that was approximately unbiased. Mixed models also have the advantage that we can do an heuristic goodness-of-fit test, as we implemented for the Carraguard data. A nonparametric test may be useful as an initial analysis in deciding whether it is worth fitting a parametric model.

If a parametric model is to be used to analyze a study of the effect of treatment on HIV RNA-1 data with one follow-up value, we recommend that it be a mixed model. We suspect that this conclusion applies more generally to data of this type, based on the simulations for cross-sectional data in Wiegand et al. (submitted for publication). Our simulation results for the crossover design suggest that using the baseline values will decrease power. Senn (2002, Section 3.16.1) comments that this is likely to occur unless the baseline values contain useful information about patients' treatment responses. Our simulation results agree with his comments, as we assumed the treatment effect to be independent of the baseline value in each period. However, it may be necessary to use baseline data in order to convince readers of the validity of analyses.

It is reasonable to believe that effective treatment will increase the variance of observed values as a result of heterogeneity of the treatment effect, as shown by the Carraguard data. As would be expected, simulation results showed that such increased variance results in substantial reduction in power for mixed models assuming equal variances and that a substantial increase in sample size is required to obtain the desired power with a model assuming unequal variances.

Our simulations to evaluate analysis methods for the crossover design when the data include a period effect found that the mixed models had substantially better power than other analysis methods. Our results also suggest that the sample size required to obtain stated power may increase when there is a period effect that reduces the baseline mean. Additional simulations with other assumptions on the period effect would be needed to make a definitive statement concerning the affect of a period effect on power.

We found that the published methods for estimating the sample size required to obtain stated power may give incorrect estimates for data with a lower quantification limit. In our simulations, Lachenbruch's method (Lachenbruch, 2001) for a parallel arm design gives an estimate that is too large when the variance of the follow-up measurements is unaffected by treatment but too small when this variance is affected by treatment. The sample size estimates for a crossover design from Senn's method (Senn, 2002) could be too large, too small, or approximately correct, depending on the proportion of non-quantifiable results. Neither method accounts for any dependence of power on the proportion of NQ values. It may be necessary to do simulations to obtain guidance on the sample size required for a treatment trial with data containing a lower quantification limit. Furthermore, our simulations assumed no loss to follow-up. Missing values could be imputed, but estimates would have larger standard errors than those from complete data.

Crossover trials are often recommended instead of parallel arm trials in order to reduce the number of persons required to obtain estimates with the same precision (Senn, 2002). For data with a lower quantification limit, our simulations confirm this recommendation if there are no period effects (fewer persons per arm or treatment sequence are needed to obtain stated power with mixed models). If there is a period effect, this advantage may be substantially reduced. Our simulations assume no loss to follow-up. While loss to follow-up would be more likely with a crossover design, experience has shown that it is much easier to retain participants in a trial than to recruit eligible participants. If a period effect would be included in a crossover study analysis model for data with a lower quantification limit, the sample size comparisons depend on the magnitude and likely the sign of the mean period effect and the standard deviation of the effect. In this case, our simulation results do not yield a clear recommendation for choosing between designs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

John Karon's work was done as a contractor for CDC. We thank two referees for comments which led to a greatly improved manuscript.

The findings and conclusions presented here are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

REFERENCES

Baeten JM, Kahle E, Lingappa JR, Coombs RW, Delany-Moretlwe S, Nakku-Joloba E, Mugo NR, Wald A, Corey L, Donnell D, Campbell MS, Mullins JI, Celum C, and the Partners in Prevention HSV/HIV Transmission Study Team. Genital HIV-1 RNA predicts risk of heterosexual HIV-1 transmission (2011). Sci Transl Med 77: 4 6;3:77ra29.

Chu H, Nie L, Cole SR (2006). Sample size and statistical power assessing the effect of interventions in the context of mixture distributions with detection limits. Statistics in Medicine 25: 2647–2657. [PubMed: 16456897]

Dunne EF, Whitehead S, Sternberg M, Thep-Amnuay S, Leelawiwat W, McNicholl JM, Sumanapun S, Tappero JW, Siriprapasiri T, Markowitz L (2008). Suppressive acyclovir therapy reduces HIV cervericovaginal shedding in HIV-1 and HSV-2-infected women, Chiang Rai, Thailand. JAIDS 49: 77–83. [PubMed: 18667923]

Farewell VT (1983). The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 3:; 1041–1046

Fedorov V, Mannino F, Zhang R (2009). Consequences of dichotomization. Pharmaceutical Statistics 8: 50–61. [PubMed: 18389492]

Gehan EA (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 52: 205–223.

Hughes JP (1999). Mixed effects models with censored data with application to HIV RNA levels. Biometrics 55: 625–639. [PubMed: 11318225]

Journot V, Chene G, Joy P, Saves M, Jacqmin-Gadda H, Molina J-M, Salamon R, and the ALBI Study Group (2001 Viral load as a primary outcome in human immunodeficiency virus trials: a review of statistical analysis methods. Controlled Clinical Trials 22: 639–658. [PubMed: 11738121]

Lachenbruch PA (2001) Comparisons of two-part models with competitors. Statistics in Medicine 20: 1215–1234. [PubMed: 11304737]

Lubin JH, Colt JS, Camann D, Davis S, Cerham JR, Seerson RK, Bernstein L, Hartge P (2004). Epidemiologic evaluation of measurement data in the presence of detection limits. Environmental Health Perspectives 112: 1691–1696. [PubMed: 15579415]

Lyles RH, Lyles CM, Taylor DJ (2000). Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. Applied Statistics 49: 485–497.

Lynn HS (2001). Maximum likelihood inference for left-censored HIV RNA data. Statistics in Medicine 20: 33–45. [PubMed: 11135346]

McLean C, van de Wijgert JHHM, Jones HE, Karon JM, McNicoll JM, Whitehead SJ, Braunstein S, Achalapong J, Chaikummao S, Tappero JW, Markowitz LE, Kilmarx PH (2010). HIV genital shedding and safety of Carraguard use by HIV-infected women: a crossover trial in Thailand. AIDS 24: 717–722. [PubMed: 20098295]

Molina J-M., Andrade-Villaneuva J, Chevarria J, Chetotisakd P, Corral J, David N, Moyle G, Mancini M, Percial L, Yang R, Thiry A, McGrath D, for the CASTLE Study Team (2008). One-daily atazanavir/ritonavir versus twice-daily lopinavir/ritonavir, each in combination with tenofovir and emtricitabine, for management of antiretroviral-naïve HIV-1-infected patients: 48 week efficacy and safety results of the CASTLE study. The Lancet 372: 646–655.

Moulton LH, Curriero FC, Barroso PF (2002). Mixture models for quantitative HIV RNA data. Statistical Methods in Medical Research 11: 317–325. [PubMed: 12197299]

Nie L, Chu H, Cheng Y, Spurney C, Nagaraju K, Chen J (2009). Marginal and conditional approaches to multivariate variables subject to limit of detection. Journal of Biopharmaceutical Statistics 19: 1151–1161. [PubMed: 20183469]

Randles RH, Wolfe DA (1979). Introduction to the Theory of Nonparametric Statistics New York: John Wiley & Sons.

Schouten H, Kester A (2010). A simple analysis of a simple crossover trial with a dichotomous outcome measure. Statistics in Medicine 29: 193–198. [PubMed: 19882677]

Senn S (2002). Cross-over Trials in Clinical Research, Second Edition. Chichester, England: John Wiley & Sons.

Senn S (2003). Disappointing dichotomies. Pharmaceutical Statistics 2: 239–240.

Senn S, Holford N, Hockey H (2012). The ghosts of departed quantities: approaches to dealing with observations below the limit of quantitation. Statistics in Medicine 31: 4380–4295. 10.1002/sim5515.

Thiebaut R, Jacqmin-Gadda H (2004). Mixed models for longitudinal left-censored repeated measures. Computer Methods and Programs in Biomedicine 74: 255–260. [PubMed: 15135576]

Thiebaut R, Jacqmin-Gadda H (2007). Corrigendom to “Mixed models for longitudinal left-censored repeated measures”. Computer Methods and Programs in Biomedicine 87: 78–79.

Wiegand RE, Rose CE, Karon JM Comparison of methods for analyzing two-group, cross-sectional data with an outcome subject to a detection limit Submitted for publication.

Table 1.

Distributions of $\log_{10}$ HIV-1 RNA (copies per mL of cervico-vaginal lavage fluid), phase I Carraguard trial, Chiang Rai, Thailand.

Group	N	% ND	% NQ	% Q	Median (Q1, Q3)
At the start of each cycle (day zero)					
Cycle 1	58	19	16	66	2.37 (NQ, 3.07)
Cycle 2	53	19	21	60	2.04 (NQ, 2.96)
Cycle 3	52	27	17	56	2.12 (ND, 2.87)
On day seven, by treatment					
Carraguard®	54	39	19	43	NQ (ND, 2.38)
Placebo gel	55	44	15	42	NQ (ND, 2.59)
No product	54	24	20	56	2.03 (NQ, 2.9)
On day 14, by treatment					
Carraguard®	51	22	24	55	2.11 (NQ, 2.73)
Placebo gel	54	37	13	50	NQ (ND, 2.95)
No product	53	25	21	55	2.13 (NQ, 3.07)

ND: none detected; NQ: detected but not quantifiable; Q: quantifiable (lower limit is 1.90 on the $\log_{10}$ scale). Q1: first quartile; Q3: third quartile.

Table 2.

Summary of the analyses of the variation of HIV-1 CVL RNA among treatments: test statistics and p-values, and treatment effect coefficients (standard error estimates in parentheses: copies/mL). Phase I Carraguard trial, Chiang Rai, Thailand.

Model	Test statistic (P-value)	Gaussian coefficients		Logistic coefficients	
		Placebo gel	Carraguard	Placebo gel	Carraguard
Nonparametric analyses: Friedman test					
NQ and ND combined	10.7 (.005)	NA	NA	NA	NA
NQ and ND separate	15.2 (<.001)	NA	NA	NA	NA
Mixed models using day seven only					
Equal variances	14.1 (0.001)	−0.439 (0.148)	−0.543 (0.150)	NA	NA
Unequal variances	14.2 (0.003)	−0.455 (0.153)	−0.561 (0.157)	NA	NA
Mixed models using both days zero and seven					
Equal variances	7.6 (0.02)	−0.526 (0.191)	−0.303 (0.193)	NA	NA
Unequal variances	7.6 (0.06)	−0.532 (0.194)	−0.309 (0.196)	NA	NA
Mixture models using the likelihood L_2 and data from day seven only					
Equal variances	4.6 (0.34)	0.036 (0.167)	−0.195 (0.171)	−0.553 (0.387)	−0.522 (0.388)
Unequal variances	5.8 (0.32)	0.036 (0.167)	−0.195 (0.171)	−0.653 (0.387)	−0.522 (0.388)
Mixture models using the likelihood L_2 and data both days zero and seven					
Equal variances	4.30 (0.37)	0.036 (0.171)	−0.195 (0.177)	−0.553 (0.362)	−0.522 (0.356)
Unequal variances	6.86 (0.23)	0.036 (0.148)	−0.195 (0.152)	−0.554 (0.363)	−0.521 (0.358)
Generalized estimating equation logistic regression models for a meaningful decrease					
NQ and ND combined	2.2 (0.34)	NA	NA	−0.345 (0.417)	− 0.550 (0.410)
NQ and ND separate	1.0 (0.62)	NA	NA	−0.152 (0.389)	− 0.338 (0.387)

NQ: sample detectable but not quantifiable; ND: sample not detectable;

NA: not applicable.

Test statistics:

Friedman test: chi-square with 2 degrees of freedom (df).

Mixed models and mixture models: likelihood ratio chi-square with 2 and 3 df for equal and unequal variances, respectively.

Generalized estimating equation logistic regression models: Wald chi-square with 2 df.

Table 3.

Power (percent) for detecting a treatment effect based on 4000 simulations in a two-arm (placebo/treatment) parallel design study, by μ and the standard deviation of the values after treatment (σ_{Trt}) Results are for at least 99% of the simulations, except as indicated.

	μ (64 per arm, σ_{Trt} = 1)				**μ (90 per arm, σ_{Trt} = 1.4)**			
	2.0	**2.5**	**3.0**	**4.0**	**2.0**	**2.5**	**3.0**	**4.0**
Model								
Nonparametric analyses								
Wilcoxon test (NQ only)	60.3	74.0	75.8	76.8	38.9	60.5	70.1	77.0
Wilcoxon test (both NQ and ND)	67.9	79.0	76.9	76.8	53.2	68.2	72.7	77.1
Gehan test (NQ only)	61.0	74.9	76.8	78.0	36.7	60.3	71.5	77.1
Mixed models using both baseline and follow-up data								
Equal variances	82.3	90.5	92.1	93.2	37.7	63.2	81.6	93.8
Unequal variances	77.0	89.9	92.4	93.1	64.2	81.4	90.8	94.5
Mixture models using likelihood L_2 and follow-up data only								
Equal variances	51.6	63.5	67.1	56.1[d]	NS	NS	NS	NS
Unequal variances	51.4	61.5	63.6	51.7[d]	NS	NS	NS	NS
Mixture models using likelihood L_2 and both baseline and follow-up data								
Equal variances	19.8[c]	25.1[b]	26.6[a]	30.6[a]	NS	NS	NS	NS
Unequal variances	28.2[c]	33.2[b]	30.7[a]	25.1[a]	NS	NS	NS	NS
Generalized estimating equation logistic regression models for a significant decrease								
NQ and ND combined	36.2	52.7	67.1	68.3	17.7	41.0	65.2	83.3
ND separate from NQ	52.5	64.5	71.4	68.8	33.5	55.2	72.0	84.0

Assumptions: treatment reduces the underlying mean μ by 0.5, the standard deviation of the untreated mean is 1, and Type I error is 0.05. Quantification limit is $\log_{10}(80) = 1.90$. NQ: non-quantifiable; ND: not detectable.

NQ only: NQ and ND were combined.

[a] Results for 90% to 98.9% of all simulations.

[b] Results for 80% to 89.9% of all simulations.

[c] Results for 70% to 79.9% of all simulations.

[d] Results for 60% to 69.9% of all simulations.

NS: not shown, excessive Type I error.

Table 4.

Probability of a meaningful decrease, power to detect a treatment effect based on this dichotomization with n=64 in each treatment group in a two-arm parallel study, and the sample size in each group to have 80% power to detect a such treatment effect, when the correlation between baseline and follow-up measurements is 0.6, by mean baseline CVL HIV-1 and mean treatment effect.

Mean treatment effect	Baseline mean								
	2.0			2.5			3.0		
	Pr(decrease)	Power (%)	N per arm	Pr(decrease)	Power (%)	N per arm	Pr(decrease)	Power (%)	N per arm
0.50	0.367	28	217	0.446	44	135	0.486	56	104
0.75	0.421	51	115	0.525	77	67	0.583	90	50
1.00	0.464	70	79	0.585	94	44	0.668	99	31

A meaningful decrease is a decrease of at least 0.5 if both measurements are quantifiable, or a decrease from quantifiable to non-quantifiable.

Quantification limit is 1.9; standard deviation of both baseline and follow-up values is 1.

The sample sizes (N per arm) were obtained from the web site http://www.swogstat.org/stat/public/binomial_twoarm.htm

Table 5.

Power (percent) for detecting a treatment effect based on 4000 simulations in a two-arm, two-treatment crossover study, by μ and the standard deviation of the values after treatment (σ_{Trt}). Results are for at least 99% of the simulations, except as indicated.

	μ (16 per arm, σ_{Trt} = 1)				**μ (30 per arm, σ_{Trt} = 1.4)**			
Model	**2.0**	**2.5**	**3.0**	**4.0**	**2.0**	**2.5**	**3.0**	**4.0**
Nonparametric analyses								
Friedman test (NQ only)	56.8	64.9	68.4	69.3	28.4	50.6	64.2	75.0
Friedman test (both NQ and ND)	61.5	66.6	70.0	68.9	41.0	57.0	66.3	75.6
Schouten/Kester test	19.5	28.6	37.4	40.7	14.4	29.4	47.3	64.7
Mixed models using follow-up data only								
Equal variances	68.9	78.4	84.6	87.1	30.5	58.4	76.8	89.9
Unequal variances	42.4[a]	70.5[a]	83.3	86.7	56.1[a]	77.2[a]	86.0	89.6[a]
Mixed models using both baseline and follow-up data								
Equal variances	44.3	51.6	56.9	58.6	18.5	36.1	53.5	71.4
Unequal variances	36.8	49.4	56.3	58.6	31.5	49.8	63.5	73.2
Mixture models: likelihood L_2, follow-up data only								
Equal variances	14.1	17.0	17.0[a]	NA	NS	NS	NS	NS
Generalized estimating equation logistic regression models for a significant decrease								
NQ and ND combined	21.0	31.1	38.9	42.5	14.3	29.4	47.3	64.7
ND separate from NQ	30.4	36.4	42.0	43.0	24.3	40.0	53.6	65.8

Assumptions : treatment reduces the underlying mean μ by 0.5, the standard deviation of the untreated mean is 1, Type I error is 0.05, the quantification limit is $\log_{10}(80) = 1.90$. NQ: non-quantifiable; ND: not detectable.

NQ only : NQ and ND results were combined.

[a] Results for 90% to 98.9% of all simulations

NA: Results for less than 20% of all simulations.

NS: not shown, excessive Type I error.

Table 6.

Power (percent) for detecting a treatment effect with a potential period effect based on 4000 simulations in a two-arm, two-treatment crossover study, by the standard deviation of the values after treatment (σ_{Trt}), and the period effect (mean μ_P, standard deviation $\sigma_{P)}$. Results are for at least 99% of the simulations, except as indicated.

	35/arm, σ_{Trt} = 1			**50/arm, σ_{Trt} = 1.4**		
Period effect: μ_P (σ_P)	**0.0 (0.0)**	**−0.3 (0.3)**	**−0.3 (0.1)**	**0.0 (0.0)**	**−0.3 (0.3)**	**−0.3 (0.1)**
Nonparametric analyses						
Friedman test (NQ only)	93.1	60.7	63.2	72.7	21.6	21.8
Friedman test (NQ and ND)	94.5	68.1	70.3	79.1	33.2	34.2
Mixed models using follow-up data only						
Equal variances	98.4	76.2	78.5	81.2	23.1	22.7
Unequal variances	96.6	62.1	65.1	94.4[a]	54.7[a]	58.0[b]
Mixed models using both baseline and follow-up data						
Equal variances	84.8	77.5	80.9	56.2	42.4	45.0
Unequal variances	83.5	74.7	78.0	74.1	62.0	65.3

Assumptions: treatment reduces the underlying mean by 0.5 from a baseline mean of 2.5, the standard deviation of the untreated mean is 1, Type I error is 0.05, quantification limit is $\log_{10}(80) = 1.90$.

NQ: non-quantifiable; ND: not detectable.

NQ only : NQ and ND results were combined.

[a] Results for 91.5% to 98.5% of all simulations

[b] Results for 89.3% of all simulations.