Designing trials to test the bioequivalence of diets for animal performance[1,2]

I. McMillan[2,3]

Department of Animal Science, University of Guelph, Guelph, Ontario, N1G 2W1 Canada

ABSTRACT: It is usually the desire of the researcher to demonstrate a difference in the comparisons made. This is certainly the case when an improved product is desired. However, in establishing the bioequivalence of a test product to a standard, the objective is usually to conclude, with reasonable justification, that no difference has been detected. In making such determinations, the probabilities of accepting false hypotheses or those of rejecting correct hypotheses of difference must be taken into account. Before beginning the trial, the researcher should have a good estimate of the power that will be associated with the detection of a given minimum acceptable difference. The required sample size for achieving the desired power for these tests depends on the coefficient of variation in the data collected and the minimum detectable difference between two groups that the study is meant to detect. It is important to determine the relative magnitudes of the sampling and the experimental errors in order to decide, for a given number of animals available for the study, how they might be best subdivided into groups within treatment. This paper addresses these points and others in an attempt to summarize some of the key items that researchers should consider when planning trials for the bioequivalence testing of new products.

Key Words: Experimental Design, Hypothesis Testing, Power, Sensitivity

©2004 American Society of Animal Science. All rights reserved. J. Anim. Sci. 2004. 82(E. Suppl.):E223–E228

Introduction

Whenever a new feed additive is developed for use in the agricultural production industry, interest naturally centers on its effects on the palatability of the diet to which it is added, and also on the performance of the animals to which it is fed compared with their performance when fed a standard diet. Some trials have been done to assess the composition and digestibility of genetically altered crops, and to compare their effects on the health and performance of livestock with what is found for the crops routinely used to produce livestock feed. Dry matter intake and BW are two commonly measured variables used in such comparisons.

The main objectives for altering plants using biotechnology have thus far been to provide herbicide tolerance and natural pest protection. The need for chemical pesticide application can therefore be reduced, and a corresponding reduction in residues found in animal feed and in food and water resources can be expected (Faust, 2002). Such results are very desirable, but the livestock producer needs to have confidence that using such a feed will not negatively affect the production of the stock that ingests it.

Product Testing

Nutritional experiments in the past for livestock have often centered on demonstrating the superiority of a new product to those already in use. The advantage of the new product is often measured as an improvement in any of a number of production traits in the stock to which it is fed. These trials are often designed to show that a statistically significant difference exists between a conventional Product A and the new Product B, and they are usually carried out in a fairly routine way. Regardless of the number of treatment groups included, the procedure normally employs an experimental design that removes identified possible bias in the comparisons to be made among the groups and also uses a design that reduces the error used to make these comparisons. The smaller that the error can be made, the smaller the difference that will appear to be significant.

Bioequivalence, on the other hand, requires a demonstrated lack of superiority of the standard Product A to the new Product B. When the intention is to show that A is not superior to B, an obviously simple approach would be to use an experimental design that does little to reduce the error to be used for testing. Although the

[1]This article was presented at the 2003 ADSA-ASAS-AMPA meeting as part of the Contemporary Issues symposium "Designing Animal Experiments for Power."

[2]Acknowlegments: Valuable comments were given by M. Quinton.

[3]Correspondence—phone: 519-824-4120, ext. 53506; fax: 519-767-0573; e-mail: imcmilla@uoguelph.ca.

Received July 2, 2003.

Accepted October 27, 2003.

bias may be removed, the sensitivity of the comparison is not improved and the conclusion of a statistically nonsignificant difference becomes more likely. It would therefore seem prudent to assess such trials for their sensitivity before accepting their conclusions.

Sample Size

For trials that are designed to compare a biotechnically altered crop to an unaltered isogenic line using, for example, the weight gains of animals fed one crop to those fed the other within a given period of time, the hypothesis of there being a difference between the two is usually tested with a statistic such as a Z or a t. When the test involves more than just a straight comparison of two group means, like the additional factor of blocking, an ANOVA is called for and the F-test becomes the initial statistic of importance. Whatever the design, the results depend on the size of the difference between average performances of the two products in the test, divided by the standard error of the difference.

To calculate the sample size required for finding a specific difference to be statistically different from zero, a formula given by Snedecor (1980) can be used:

$$n = 2(Z_\alpha + Z_\beta)^2\left(\frac{\sigma}{d}\right)^2 \quad [1]$$

where n is the sample size required to show a difference of $d = \overline{x_A} - \overline{x_B}$ units between the means of the standard Product A and the new Product B to be significant at the α level. The Z characters are values for the normal deviate at the α and β levels of significance, and σ is the standard deviation among the experimental units within a group. If only the absolute value of the difference between the standard and new means is of interest and there is a specific absolute value below which the difference would not be considered to be of importance, then the hypothesis to test is two-tailed and appears as Ho: $d = d_0$. The alternate hypothesis, Ha: $d \neq d_0$, is also two-tailed. If the intention is to test whether the new product is at least as good as the standard, the null hypothesis is stated as Ho: $d \leq d_0$, where d_0 is the minimum difference between the groups that is to be detected. The test is now one-tailed. The alternate hypothesis, Ho: $d > d_0$, is also one-tailed. In both the one- and two-tailed situations, the probability of rejecting a true null hypothesis and accepting a false alternate hypothesis is α, and β is the probability of accepting a false null hypothesis rejecting a true alternate. Values of d less than d_0 are not, on average, detectable using the trial design and size under consideration. The magnitude of d_0 is commonly chosen with a biological and, sometimes, an economic criterion in mind. Anything below the given value is not considered to be detrimental to the performance of the new product. Berndtson (1991) refers to d_0 as the "sensitivity" of the test.

The power of the test is given by $1 - \beta$, the chance of rejecting a false null hypothesis. In the present case, it represents the chance of rejecting the bioequivalence of the test product, for a particular variable, to the standard when the difference is equal to or greater than d_0. The calculation of the power of a statistical test is a very important tool in assessing the confidence one can place in the results of a comparison. Almost any area of biological research provides logical circumstances for doing so. Ling and Cotter (2003) provided formulae relating to the calculation of power in aquaculture growth trials that use a nested design. They calculated the minimum number of tanks and fish in order to arrive at the most cost-effective design for comparative growth trials. McMillan and Robertson (1974) described the chances of detecting the presence of a locus in gene-mapping trials. The technique investigates the heterogeneity in the expression of a trait among the resulting chromosome groups formed in a crossover trial. The probabilities of finding group differences depend on both the size and the standard deviation of the gene effect which are used in calculating the power of the trial design.

Coefficient of Variation

The coefficient of variation of a variable is a measure of its standard deviation expressed as a fraction of the sample mean (m):

$$CV = \left(\frac{\sigma}{m}\right) \quad [2]$$

It is often used to provide an estimate of how many observations would be required for each of the test and control groups to show a specific difference between them as being statistically significant for a given variable. For stated α and β error rates that pertain to the null and alternative hypotheses being tested, the CV allows a calculation of the minimum difference, as a proportion of the control group mean, that can be shown to be statistically significant. Dividing both the numerator and denominator in Eq. [1] by the mean of the control group, m, gives:

$$n = 2(Z_\alpha + Z_\beta)^2(CV)^2/(d/m)^2 \quad [3]$$

The CV and the d/m ratio can be each expressed as a percent. Using this formula, tables can be constructed to examine the effect of changing the CV percentage on the magnitude of the difference between groups that can be detected for a given sample size. Table 1 shows what sample size would be required to find differences of 5, 10, and 20% significant for CV percentages of 5, 10, and 20. The values for $1-\beta$ are 0.8 and 0.9 and each calculation is done for a one-tailed test with α at the 0.05 and 0.025 levels of significance.

It is apparent from Table 1 that the sample size required to detect a given difference increases as the difference becomes smaller, and also when the power of the test increases. It is also obvious that, for a given

Table 1. Sample sizes (n) required to show a difference between two groups to be significant for various levels of α and β using a one-tailed test[a]

1-β	CV, %	Difference, %[a]	n for α = 05	n for α = 0.025
0.80	5	5	13	16
0.80	5	10	4	4
0.80	5	20	2	2
0.80	10	5	50	63
0.80	10	10	13	16
0.80	10	20	4	4
0.80	20	5	198	252
0.80	20	10	50	63
0.80	20	20	13	16
0.90	5	5	18	22
0.90	5	10	5	6
0.90	5	20	2	2
0.90	10	5	69	85
0.90	10	10	18	22
0.90	10	20	5	6
0.90	20	5	275	337
0.90	20	10	69	84
0.90	20	20	18	21

[a]The difference is expressed as a percentage of the standard group mean.

difference and power, an increase in the CV percentage requires an increase in the sample size.

The use of Z in Eq. [1] assumes that σ^2 is known, rather than approximated. An approximation of σ^2 means that a t-statistic is more properly used for comparing the group means. However, because the value of the critical t-value increases as the degrees of freedom decrease, a sequential approach is required in which the calculation of n is repeated until an acceptable solution is found that works for the design of the trial (Steel et al., 1997). It is not unusual to employ the Z value when σ^2 is estimated with a reasonable number of degrees of freedom. Berndston (1991) provided tables based on the t-statistic for samples sizes required to detect specific differences for various levels of α and 1 – β. Steele et al. (1997) suggested that the sample size calculated using Z values would be increased by a factor of (Error df +3)/(Error df+1) when t-values are used instead. Error df comes from the ANOVA table used for the analysis. When the degrees of freedom for the error term are not large, this factor can be quite large. For example, if the trial is performed with a nested component, such as would be the case for dairy cattle kept in pens that were given either the test diet or the control diet, the proper error for testing the diet effect would be given by the pen within diet mean square. Even if there are several animals in each, the pens are the experimental units upon which the error is based. So if each pen had five animals, but only two pens were used for each diet, the degrees of freedom for the error would be two. The factor converting the required sample size calculated with Z values to one using t-values would be 5/3 and the trial would be increased in size by more than 60%. When planning such a trial that will have a nested effect, which represents the proper error to use for testing diet effects, the numbers of pens to include depends primarily on the CV that the pens within diet will give and also on the sensitivity required of the comparison.

When the pen-within-diet effect is not significant, it is tempting to use the sampling error with 16 df to test diets. The statistical model constructed for the trial has included the pen within diet effect. If this effect has only two degrees of freedom, testing its significance is not done with very much power. Nevertheless, it is the best estimate of the error variance and should be used as such, assuming that the pen within diet effect is random. If the pen effect is considered fixed, then it is not a true error term and the results of the study apply to only the pens used in the trial. If more than one farm is used, a confounding of farm and pen effects is created, and it would be more appropriate to analyze the data on each farm separately.

Some care must be taken in choosing the levels of α and β. The researcher is interested in accepting the hypothesis that the new product is at least as good as the standard for the variable under test. The difference, $d = \bar{x}_S - \bar{x}_N$, between the means of the two products is therefore tested by the null hypothesis, Ho: $d \le 0$. The alternative hypothesis, indicating that the new product does not perform as well as the standard would be Ha: $d > 0$. This is a one-tailed test of significance. If the researcher wanted to show that the new product is superior to the standard, the hypotheses would appear as Ho: $d \ge 0$ and Ha: $d < 0$. Again, this is a one-tailed test of significance. A two-tailed test where Ho: $d = 0$ and Ha: $d \ne 0$ asks whether the new product is equal to the control group or whether it is different, either better or worse. The two-tailed test may have a place in bioequivalency trials, but the researcher must be clear on what is being tested, and why.

For a one-tailed test, Z_α is the term to use in Eq. [1]. If the test is two-tailed, $Z_{\alpha/2}$ must be used. As can be

seen from Table 1, the sample sizes required for finding a percentage difference to be significant when α is 0.05 are less than those required when α is 0.025, so a one-tailed test would find smaller differences significant than would a two-tailed test.

Considerations Relating to the Experimental Design

To show a stronger bioequivalence, it is necessary to find, with a reasonable power, a smaller difference between the biotechnically altered group and the unaltered isogenic line to be nonsignificant. Returning to Eq. [1], it is obvious that the magnitude of the error, σ^2, is a critical component in deciding how small the value of d is that can be detected. But, it is important that the appropriate σ^2 be used. This depends on the design of the trial and the intended application of the results. For example, if the comparison of the test and control products is made with data collected from a feeding trial on one farm, then a simple t-test can be done and the results apply only to that farm. If two breeds of animal are also included, then an ANOVA can be done as a 2×2 factorial to see if any interaction between the breeds and the products exists. This is a test to see whether the difference between the products is the same for each breed. If it is not, then the products must be compared with each other separately, within each breed, and the sample sizes for these tests will be correspondingly reduced. For a balanced design, the reduction will be 50%. Again, the results of such a trial apply only to the farm on which the trial was performed. If the intention is to have the results apply to more than one farm, a second, third, or however many farms are desired, can be used to perform the trial simultaneously. The F-test in ANOVA is one-tailed since the objective is to show that the variance between treatment groups is larger than the variance within the groups. It is the equivalent of a two-tailed t-test since the sign of the difference does not affect the significance of the calculated F-statistic. If the comparison is meant to be one-way, then the sign of the difference must be noted.

If only these farms are of interest, then the analysis can be done as a randomized complete block design with farms as the blocks. In this situation, both farm and product are considered to be fixed effects. However, if the farms are considered as a random sample of all the farms that the results should be applicable to, then these farms in the trial are a random effect. The products are still a fixed effect because they are the only products of interest, so the model for the analysis is a mixed effects model. The expected mean squares and degrees of freedom for such a case are given in Table 2. These values are calculated assuming that the pen totals, and not the individual animal measures, are used. It is the pen that is the experimental unit, of which the within-pen measures are subsamples. A more refined analysis to estimate the subsampling error can be performed, but becomes tedious to do by hand, especially when the data are unbalanced. Using a statistical computer package such as SAS (SAS Inst., Inc., Cary, NC) becomes necessary for such cases.

Table 2. Degrees of freedom and expected mean squares for a randomized complete block mixed model analysis to compare two diets, test and control, administered to animals on more than one farm[a]

Source	df	Expected mean square
Total	2rp − 1	
Farm	r − 1	$\sigma^2 + p\sigma_\varepsilon^2 + 2p\sigma_F^2$
Diet	1	$\sigma^2 + p\sigma_\varepsilon^2 + rp\theta_D^2$
Residual	(r − 1)	$\sigma^2 + p\sigma_\varepsilon^2$
Sampling error	r(p − 1)	σ^2

[a]The diet effect is fixed, and the farm effect is random. The analysis is performed on pen totals or means and assumes the data are balanced. r = number of farms; p = number of pens per treatment per farm; σ_F^2 = variance component for random farm effect; θ_D^2 = variance component for fixed diet effect; and σ_ε^2 = variance component between farms.

Because the comparison of the two diets is the main objective of the trial, and since the feeds given to the animals are usually from one source with little or no replication, the individual animals within each of these groups are sampling units of the diets used. They allow for the calculation of a sampling error to measure the variation among the animals on either of the two diets within each farm. The experimental error that is used for testing the hypotheses of interest is based on the farm-to-farm variation, as well as the within-farm variance. When only two farms are used, the experimental error will have only one degree of freedom and the minimum difference that can be detected will be quite large. Conclusions relating to bioequivalence in such a situation appear a bit meaningless. If the experimenter tries to increase degrees of freedom by repeating the trial, the correct error to use for the diet comparison depends very much on whether the same two farms are used, which would result in a split plot in time, or on whether different farms are used. The latter situation merely repeats the randomized complete block design set up with a marginal increase in the degrees of freedom for error. The split in time is properly done as a repeated measures analysis and if there is an interaction of product and time then the diet comparisons must be done separately within each time period with one degree of freedom for error in each case.

The review paper by Clark et al. (2001) includes data that enable the calculation of the CV for a variety of variables in the different species covered. One study described in this paper concerns a comparison of Bt corn with a control non-Bt isoline fed to chickens from one to 38 d of age (Brake and Vlachos, 1998). This product is recognized as having pest control action. The trial measured final BW, feed gain, percentage alive, and several carcass components weights expressed as a percentage of BW. The sample sizes were 640 birds

per treatment and the CV ranged from a low of 12.3% for the carcass component wings to 61.8% for BW. Another study described in this review was set up to compare early maturing versus late maturing corn silages, again of Bt corn vs. a non-Bt isoline (Folmer et al, 2000). The variables used included DMI, ADG, and G:F. Their CV were 7.4, 12.5, and 10.0%, respectively. The sample size for each of the four treatment groups in this study was 32. These studies concluded that no significant difference was found between the bio-technically altered crop and the non-Bt isoline.

Looking at the values in Table 1, it is obvious for a given level of power that the sample size required to detect a minimum difference of 10% based on a CV of 10% for one of the variable may not be large enough for a variable with a CV of 20%. This leads to a consideration of how to choose an appropriate CV for calculating the sample size for the experiment. The process might be easily done for trials where one variable is more important than any of the others. Body weight in fish, for example, is usually the main trait of interest, whereas carcass quality traits are of secondary importance. But when it is difficult to identify one trait to be more important to the production value of the animal than two or three of the others, the task is not so easily performed, especially when the CV differ substantially among them. One must then take into account the time and the facilities available for conducting the trial and decide on whether using the highest CV of the variables of interest gives a required sample size that can be handled. If it is too large, then the next lowest CV of these variables might give a workable sample size. If not, then one should continue down to the next lowest CV. It should be noted that for those variables with CV lower than the one used to calculate the sample sizes, the test will be slightly more sensitive and smaller differences between the diet groups will turn out to be significant than for the variable used to calculate the required n.

Other Considerations

When there are two or more production variables of interest, the procedure for testing the bioequivalency of the new product can be done either as a series of univariate tests that consider each variable separately, or as a multivariate analysis that considers all of the variables at once. This latter procedure is not commonly used in the testing of bioequivalency since the significance of a single, economically important trait can be masked in the average significance calculated for the whole group of traits under consideration. A common correction is to test the separate variables using the desired experiment-wise error divided by the number of tests to be done. Some prefer using the square root of the advisor in this calculation.

When only two groups are being compared and the analysis is done as a series of univariate *t*-tests on the variables of interest, the chance that one of them will appear significant when none is, is the experiment-wise error rate. The α used for the separate tests is the comparison-wise error rate. When each of the comparisons is performed at a given level of α, the overall chance of rejecting at least one of them is increased to a level higher than α. If the tests are independent and t of them are performed the experiment-wise error can be calculated as $1-(1-\alpha)^t$. A Bonferroni correction (Kuehl, 2001) could be applied to determine the proper value for α to use so that the experiment-wise error rate is kept to an acceptable level. Table 3 lists the unadjusted experiment-wise error rate and the adjusted comparison-wise error rates for different numbers of tests preformed. As the number of tests done increases, the chance of at least one significant result increases when no true difference exists. To keep the experiment-wise error rate at a fixed value of 0.05, the comparison-wise error rate becomes quite small and the minimum difference that can be detected increases, thereby reducing the sensitivity of the test.

Table 3. Changes in the comparison wise error rate to keep the experiment-wise error at 0.05 when making t independent comparisons using a comparison-wise error of α/t (Bonferroni inequality)[a]

No. of tests (t)	Experiment-wise error (unadjusted)	Comparison-wise error (adjusted)
1	0.05	0.05
5	0.22	0.01
10	0.40	0.005
20	0.64	0.0025

[a]Unadjusted experiment-wise = $1 - (1 - \alpha)^t$; adjusted comparison-wise = α/t.

Chesson and Flachowsky (2003) noted that the introduction of new genetic material into a plant has only a remote possibility of having unintended effects that would not be detected by chemical analysis. They make the observation that since broilers are sensitive to any change in the nutrient content or presence of toxic elements in their feed because of their rapid weight gain, their rate of growth could be the trait to consider when measuring genetically modified products for unintended effects. In order to cut down on the number of tests performed in an analysis, it would be of value to consider and limit to a reasonable number those traits that could be used as a standard measure to judge the safety of the test product. This would give more reasonable experiment- and comparison-wise error rates than those often seen in the current literature.

One detail that is often overlooked in the analysis is the identification of the proper experimental unit that gives rise to the variance used for testing the difference of interest. For example, whenever animals are housed in groups, such as chickens or pigs in pens, it is the group in the pen that is often the experimental unit, not the animal within the pen, because feeding trials are usually conducted by applying the diet to the pen,

not to the animals individually. Therefore, the degrees of freedom for the error term can be overestimated if the wrong experimental unit is taken. Also, variance between pens can be larger than variance within pens and the error term must include both.

When the variable to be measured is a single event, such as weight gain at a given time, the analysis is usually straightforward. However, when repeated observations are made over time on the same animals with the intention, for example, of comparing weight gain at various times during the trial, the correlation between weights at two different times does not necessarily remain constant for all possible pairs of times at which the weights are recorded. This results in a lack of compound symmetry, a term that refers to a situation in which a repeated-measures ANOVA must be done (Kuehl, 2001). This sets up a split-plot in time construct for the analysis in which the subplot error is broken up into the appropriate parts for testing elements of the (time × treatment) interaction. Each of the error components used for these tests has fewer degrees of freedom than the complete subplot error term. If any part of the interaction is significant, the group comparisons must be made at each time separately. The degrees of freedom for error are thus reduced and an increase in the magnitude of the minimum difference that can be detected is therefore required. A detailed discussion of repeated-measures analysis and techniques to perform them is given by Littell et al. (1998).

Food safety, both for the animals eating genetically altered crops as well as for the humans consuming animal products, is commonly measured as a frequency of occurrence of illness or death. Proportional data usually are not normally distributed variables. More often, they are of binomial form and the variance of such a variable is proportional in some way to its mean value. Consequently, if there is a difference among the group means within a trial their variances will likely be heterogeneous and a proper analysis of variance will require a prior transformation of the data to remove the variance to mean relationship. This is especially the case when the proportions lie outside of the 0.3 to 0.7 range (Steel et al., 1997). The arcsin transformation is often recommended for this situation, under the assumption that the sample sizes upon which the proportions are based are all equal. When they are not, an alternative transformation to use is the logit.

The χ^2 is a useful statistic to use for comparing the incidences of illness or mortality between groups. When there are more than two groups, the reason for any significance of a calculated χ^2 is sometimes not obvious. When only two groups are involved, then it is usually apparent which group has the higher proportion. However, when the group sizes upon which the proportions are based fall below five, then the χ^2 is not a reliable statistic to use and a Fisher's exact probability calculation is more appropriate (Steel et al., 1997).

Implications

When performing trials that compare the production potential of a new product with what is presently available in an existing product, it is imperative to specify the minimum difference between the two that will be used as a criterion for bioequivalence. This difference should have biological significance. It is also essential that the appropriate design be chosen to control bias and make possible a representative sample of the effects the products deliver in the experimental units that they are applied to. Using an incorrect error term for making the required comparisons is an often overlooked item that can result in misleading results. It is important that the researcher be quite clear on how broadly applied the results of the trial are intended to be. When there is concern about the validity of the statistical model or the way the analysis is to be performed, a statistician should be consulted. It is wise to involve a statistician in the planning stage of the trial.

Literature Cited

Berndston, W. E. 1991. A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments. J. Anim. Sci. 69:67–76.

Brake, J., and D. Vlachos. 1998. Evaluation of transgenic event 176 'Bt' corn in broiler chickens. Poult. Sci. 77:648–653.

Clark, J. H., and I. R. Ipharraguerre. 2001. Livestock performance: Feeding biotech crops. J. Dairy Sci. 84:(E Suppl.):E9–E18.

Chesson, A., and G. Flachowsky. 2003. Transgenic plants in poultry nutrition. World Poult. Sci. J. 59:201–207.

Faust, M. A. 2002. New feeds from genetically modified lants: The US approach to safety for animals and the food chain. Livest. Prod. Sci. 74:239–254.

Folmer, J. D., C. E. Erickson, C. T. Milton, T. J. Klopfenstein, and J. F. Beck. 2000. Utilization of Bt corn residue and corn silage for growing beef steers. J. Anim. Sci. 78(Suppl. 2):85. (Abstr.)

Kuehl, R. O. 2001. Statistical Principles of Research Design and Analysis. Wadsworth, Inc., Belmont, CA.

Ling, E. N., and D. Cotter. 2003. Statistical power in comparative aquaculture studies. Aquaculture 224:159–168.

Littell, R. C., P. R. Henry, and C. B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. J. Anim. Sci. 76:1216–1231.

McMillan, I., and A. Robertson. 1974. The power of methods for the detection of major genes affecting quantitative characters. Heredity 32:349–356.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical Methods. 6th ed. The Iowa State Univ. Press, Ames.

Steele, R. G. D., J. H. Torrie, and D. A. Dickey. 1997. Principles and Procedures of a Biometrical Approach. 3rd ed. McGraw Hill, New York, NY.