Applying Bailer's Method for AUC Confidence Intervals to Sparse Sampling

Jerry R. Nedelman,[1,3] Ekaterina Gibiansky,[1] and David T. W. Lau[2]

Received March 11, 1994; accepted July 26, 1994

Bailer (1) developed a method for constructing confidence intervals for areas under the concentration-vs-time curve (AUC's) with only one sample per subject but with multiple subjects sampled at each of several time points post dose. We have modified this method to account for estimation of the variances. How the need to estimate variances affects study design is discussed. An extension of Bailer's method is proposed where variances are modeled as a function of the means, in order to get more precise estimates of variances. The modified and extended methods are applied to a rat toxicokinetic study with only two rats per time point per treatment group.

KEY WORDS: area under the curve; degrees of freedom; Satterthwaite's approximation; variance function; toxicokinetics.

INTRODUCTION

Bailer (1) described a technique for estimating the mean area under the curve (AUC) of drug concentration vs time (CxT) when only one sample per subject is available but with multiple subjects sampled at each of several time points post dose. He also demonstrated how to estimate the standard error of the estimated AUC and how to test statistically for the equality of two mean AUC's or construct a confidence interval for the difference of two mean AUC's.

Bailer's method is elegant and simple. The tests and confidence intervals do not depend on any models for pharmacokinetic response, nor on any assumptions about variance homogeneity. They depend on the assumption of normally distributed data, and as applied by Bailer, they assume that the variances are known.

Whereas Bailer considered AUC's computed between two finite time points, Yuan (2) showed how to extend his method to infinite time, provided one has an estimate of terminal elimination rate.

We are interested in applying Bailer's method, without Yuan's extension, to rodent toxicity studies in drug development (3). In such studies, the toxicokinetic objective is to quantify drug exposure and relate exposure to dose, sex, and duration of dosing. Typically, a treatment group consists of ten or fewer animals, with only one blood sample collected per animal per day. By sampling blood from different animals at different times post dose, the CxT curve can be characterized. Bailer's method can be applied if two or more animals within each treatment group are sampled at each time point. With only ten animals per group, however, replicating at each time point severely restricts the number of time points. We typically use a design with two animals at each of five time points.

This report is about the utility of Bailer's method for such a sparse design. Our investigation revealed a complication of Bailer's method that has minor consequences for the sort of data to which the method has been previously applied (1,2), with 3-4 animals sampled per time point, but which can have more serious consequences when only two animals are sampled per time point. This complication relates to the choice of critical values used in the statistical procedures for hypothesis testing and confidence-interval estimation. That choice is related to the assumption of known variances. In this report we will focus on confidence intervals. In addition to intervals for differences between AUC's, we are also interested in intervals for single AUC's. Such intervals are useful not only to confirm exposure in the individual treatment groups, but also to provide a basis for comparing exposures in the toxicology species with exposures in later human trials.

METHODS

Suppose a study involves J treatment groups in which measurements will be taken at K time points t_k, k = 1, . . . , K; and that at time point t_k, blood is sampled from r_k animals in each group. Let u_{jkl} be the measured drug concentration from the l'th animal at time t_k in the j'th group. Let $\bar{u}_{jk}$ and s^2_{jk} be the sample average and sample variance from the r_k replicates at time t_k. Let μ_{jk} and σ^2_{jk} be the population mean and population variance of which $\bar{u}_{jk}$ and s^2_{jk} are estimates.

Bailer's method is to estimate the mean AUC for the j'th group by applying the trapezoidal rule to the $\bar{u}_{jk}$'s:

$$\hat{AUC}_j = \sum_{k=1}^{K} w_k \bar{u}_{jk} \tag{1}$$

where the trapezoidal weights, w_k, are

$$w_1 = (t_2 - t_1)/2 \tag{2a}$$

$$w_k = (t_{k+1} - t_{k-1})/2 \tag{2b}$$

$$w_K = (t_K - t_{K-1})/2 \tag{2c}$$

The variance of $\hat{AUC}_j$ is

$$\sigma^2(\hat{AUC}_j) = \sum_{k=1}^{K} w_k^2 \sigma^2_{jk}/r_k \tag{3a}$$

an estimator of which is

$$s^2(\hat{AUC}_j) = \sum_{k=1}^{K} w_k^2 s^2_{jk}/r_k \tag{3b}$$

[1] Departments of Clinical Pharmacology, Drug Safety, Sandoz Research Institute, Sandoz Pharmaceuticals Corporation, 59 Route 10, East Hanover, New Jersey 07936.

[2] Department of Drug Metabolism and Pharmacokinetics, Drug Safety, Sandoz Research Institute, Sandoz Pharmaceuticals Corporation, 59 Route 10, East Hanover, New Jersey 07936.

[3] To whom correspondence should be addressed.

0724-8741/95/0100-0124$07.50/0 © 1995 Plenum Publishing Corporation

To construct a confidence interval for the difference of two mean AUC's, say AUC_1 and AUC_2, Bailer assumed that $s^2(\hat{AUC}_1) + s^2(\hat{AUC}_2)$ is in fact the true variance of the difference $\hat{AUC}_1 - \hat{AUC}_2$, rather than just an estimate of that variance. The resulting confidence interval was of the form

$$\hat{AUC}_1 - \hat{AUC}_2 \pm z_{crit}\sqrt{s^2(\hat{AUC}_1) + s^2(\hat{AUC}_2)} \quad (4)$$

with a critical value, z_{crit}, from a standard normal distribution. The use of such a critical value rather than a critical value derived from a t distribution represents the assumption that the square-root's argument is a known rather than an estimated variance. A t_{crit} would be larger than z_{crit}, making the confidence interval wider to reflect the uncertainty in the variance.

Although Bailer did not discuss confidence intervals for a single mean AUC, we will include under the name "Bailer's method" not only intervals such as (4), but also intervals of the form

$$\hat{AUC}_j \pm z_{crit}\sqrt{s^2(\hat{AUC}_j)} \quad (5)$$

for a single mean AUC.

Generally, substituting sample variances for population variances is safe when sample sizes are large enough, for then t_{crit} approximates z_{crit}. Here, however, the adequacy of assuming that $s^2(\hat{AUC}_j)$ is in fact $\sigma^2(\hat{AUC}_j)$ in (4) and (5) depends not only on the r_k's (i.e., on the sample sizes), but also on the w_k's and the σ^2_{jk}'s. This is because $s^2(\hat{AUC}_j)$, as a weighted sum of sample variances, has a complicated distribution that can be approximated as a chi-square with degrees of freedom (df), ν_j, given by Satterthwaite's approximation (4):

$$\nu_j = \left(\sum_{k=1}^{K} w_k^2\sigma_{jk}^2/r_k\right)^2 \Big/ \sum_{k=1}^{K} (w_k^4\sigma_{jk}^4/[r_k^2(r_k - 1)]) \quad (6a)$$

It can be demonstrated that

$$\min(r_k - 1,\ k = 1, \ldots, K) \leq \nu_j \leq \sum_{k=1}^{K} (r_k - 1) \quad (7)$$

Moreover, for $s^2(\hat{AUC}_1) + s^2(\hat{AUC}_2)$ the approximating chi-square distribution has a Satterthwaite df of a similar form,

$$\nu = \left(\sum_{j=1}^{2}\sum_{k=1}^{K} w_k^2\sigma_{jk}^2 \Big/ r_k\right)^2 \Big/ \sum_{j=1}^{2}\sum_{k=1}^{K} (w_k^4\sigma_{jk}^4/[r_k^2(r_k - 1)]) \quad (6b)$$

The result can be less than the sum of the two separate df's.

By the "Bailer-Satterthwaite method" we will mean constructing confidence intervals as in (4) – (5) with z_{crit}, the critical value assuming known variances, replaced by t_{crit}, a critical value based on the t-distribution with the Satterthwaite df.

If prior information about the σ^2_{jk}'s is available, the experiment could be designed to improve the chances for a larger df. The form of (6) indicates that more replicates

Table I. Nominal 95% Confidence Intervals for Published Data

	Bailer's Data[a]						
	Bailer's Method			Bailer-Satterthwaite Method			
Dose[b]	Lower[c]	Upper[c]	Coverage[d]	Lower	Upper	DF	Coverage
LOW	.036	.062	87.4	.0319	.066	5.4	94.1
MID	.269	.455	89.7	.255	.469	9.3	94.7
HI	.506	.631	89.9	.477	.660	3.7	95.1
MID-LOW	.063	.250	89.6	.049	.264	9.7	94.7
HI-MID	.196	.323	89.4	.169	.350	4.0	95.0
HI-LOW	−.009	.215	91.6	−.020	.227	12.9	95.1
	Yuan's Data[e]						
	Bailer's Method			Bailer-Satterthwaite Method			
Sex[f]	Lower	Upper	Coverage	Lower	Upper	DF	Coverage
Female	898	1010	89.1	882	1026	5.2	94.9
Male	798	1046	91.2	738	1105	3.6	95.7
M-F	−152	120	91.8	−194	161	5.0	95.2

[a] Pre-phenylmercapturic acid in the livers of mice, μmole/g.
[b] LOW, MID, or HI are dose rates. MID-LOW, HI-MID, and HI-LOW are differences between the respective dose rates.
[c] "Lower" and "Upper" are endpoints of 95% confidence intervals computed for the actual data reported in Bailer's and Yuan's publications.
[d] Coverage is the percent coverage from the Monte Carlo simulation described in the text.
[e] Plasma pentachlorophenol concentrations after gavage administration of pentachloroanisole to B3C3F1 mice, μg/mL.
[f] M-F is MALE-FEMALE.

Table II. Blood Concentrations (ng/mL) of CPI 975 in Rats

Time post dose (hr)	Dose (mg/kg)					
	10		30		100	
	Sex		Sex		Sex	
	F	M	F	M	F	M
1	0.00	84.90	353.00	391.00	2790.00	1910.00
	126.00	136.00	384.00	396.00	3280.00	2550.00
2	128.00	194.00	625.00	649.00	4980.00	4230.00
	194.00	198.00	1410.00	1990.00	7550.00	5110.00
4	378.00	338.00	1020.00	3290.00	5500.00	7490.00
	1060.00	489.00	1500.00	3820.00	6650.00	13500.00
8	138.00	298.00	933.00	844.00	2250.00	4380.00
	146.00	a	1030.00	1650.00	3220.00	5380.00
24	0.00	0.00	0.00	75.70	213.00	260.00
	0.00	0.00	80.50	288.00	636.00	326.00

[a] Insufficient sample for analysis.

should be sampled, i.e., r_k should be larger, where $w_k^2\sigma^2_{jk}$ is larger. Prior information need not be full knowledge of the σ^2_{jk}'s. Because the w_k^2's can vary greatly (e.g., see Table IV), crude estimates of the σ^2_{jk}'s may suffice to order the $w_k^2\sigma^2_{jk}$'s.

Or, it might be possible to estimate the variance more precisely if something is known about how the σ^2_{jk}'s change with μ_{jk}. Suppose, for example, it is known that measured concentrations have a constant coefficient of variation; i.e., $\sigma_{jk} = c\mu_{jk}$, for some c. By fitting such a model to the data from all treatment groups together, and using the fitted value for the σ^2_{jk} instead of s^2_{jk} in (3), the use of z_{crit} in (4)-(5) may be a more adequate approximation, even though the variances are still not known. We will refer to such an approach as the "Bailer-model method".

The relative performances of Bailer's method, the Bailer-Satterthwaite method, and the Bailer-model method were assessed by Monte Carlo simulation. All such computer experiments were run on VAX 4000-90 workstations using SAS 6.08 under VMS 5.5-2.

RESULTS

Application to Published Data

Bailer (1) and Yuan (2) both presented data to serve as examples. In Bailer's data, there were $r_k = 4$ replicates at each of $K = 5$ time points; in Yuan's data, $r_k = 3$ and $K = 9$. Table I shows confidence intervals for each single treatment group and for the pairwise differences from those examples, using both Bailer's method and the Bailer-Satterthwaite method. The latter intervals are wider than the former. This is to be expected, since assuming that the variances are known results in the appearance of greater precision. However, the appearance can be deceiving. Table I also shows the results of a simulation study where the two methods were compared for coverage. In repeated experiments, 95% confidence intervals should cover the true value 95% of the time. Using observed sample mans as true values, the experiments were replicated 1000 times by Monte Carlo simulation assuming normal distributions (truncated at zero) and constant coefficient of variation. From Table I it is evident that the known-variance assumption of Bailer's method produces confidence intervals with only 90% coverage instead of the nominal 95%, whereas the Bailer-Satterthwaite method achieves nominal coverage. When simulations were run assuming the same means and variances but using only two replicates per time point instead of Bailer's four or Yuan's three, the coverage of Bailer's method was sometimes as low as 80%; the Bailer-Satterthwaite method maintained nominal coverage, but intervals were up to five times wider than those obtained using Bailer's method.

Application to a Rat Toxicity Study

For the compound CPI 975 under development at Sandoz, a four-week toxicity study in rats was conducted as described in the Introduction. There were six treatment groups determined by two sexes and three once-daily, gavage dose levels (10, 30, and 100 mg/kg/day). Each group

Table III. 95% Confidence Intervals for AUC's of CPI[a]

Group[b]	Bailer-Satterthwaite			Bailer-Model	
	Lower	Upper	DF	Lower	Upper
LOW F	−8889	16589	1.01	2029	5671
LOW M	1731	7407	1.01	2294	6845
MID F	12248	19006	3.19	8163	23092
MID M	−5469	59001	1.30	15120	38412
HI F	33477	86294	1.87	35751	84021
HI M	32680	148817	1.58	49650	131848
MID-LOW	−40.22	611.4	2.49	12.84	558.3
HI-LOW	94.07	570.3	3.33	52.87	611.5
HI-MID	−245.6	338.8	3.19	−285.0	378.2

[a] ng * h/mL, for single treatments. For differences between treatments, dose-normalized blood concentrations were used, so units are (ng * h/mL)/(mg/kg).

[b] LOW, MID, or HI doses are 10, 30, and 100 mg/kg. F and M are Female and Male. MID-LOW, HI-LOW, and HI-MID are differences between the dose-normalized concentrations of the respective dose levels, averaging over sex.

Table IV. Components of Trapezoidal-Rule Calculations for CPI Blood-Concentration Data

	Times Post Dose, t_k (h)				
	1	2	4	8	24
$\bar{u}_k$[a]	15.90	37.16	73.24	32.41	2.43
w_k	0.5	1.5	3.	10.	8.
w_k^2	0.25	2.25	9.	100.	64.
s^2_k[b]	20.91	288.70	765.97	77.25	6.30
$w_k^2 s^2_k/2$	2.61	324.79	3446.86	3862.51	201.62

[a] Dose-normalized concentrations averaged over all six dose-sex groups.

[b] Sample variances of dose-normalized concentrations averaged over all six dose-sex groups.

consisted of 10 rats. From each rat, blood samples were collected following drug administration on day 1 and on day 22. Two rats from each group were sampled at 1, 2, 4, 8, and 24 hours post dose. We will refer to this design, with two rats at each of five post-dose time points, as 2-2-2-2-2. Only the data from day 1 will be considered here. Table II displays that data.

Table III displays estimated confidence intervals for the CPI data using the Bailer-Satterthwaite method. (The missing value for the low-dose male at 8 hours was assumed equal to the other replicate's value of 298 ng/mL.) The estimated degrees of freedom are low for the single AUC's, and the intervals are wide. Indeed, some confidence intervals for single AUC's have negative lower endpoints. (The final two columns of Table III are discussed later.)

Table IV displays the trapezoidal weights w_j^2 and the sample averages and sample variances of dose-normalized blood concentrations averaged over all six groups. From (6) and the final row of Table IV, it is evident that only two time points, 4 and 8 hours post dose, contribute nonnegligibly to the Satterthwaite df. Were this known in advance, more replicates could have been assigned to those time points. Indeed, the first and last time points contribute so little that one replicate from each of those time points could be transferred to the important points, changing the design from 2-2-2-2-2 to 1-2-3-3-1. In that scenario, the variances at the first and last time points could not be estimated; but since those time points make such a negligible contribution, the sample variances there could be assumed to be zero.

Table V displays the results of a Monte Carlo simulation to compare the 2-2-2-2-2 design to the 1-2-3-3-1 design. Using the real data's observed sample means as true values, the experiments were replicated 1000 times by Monte Carlo simulation assuming normal distributions (truncated at zero) and constant coefficient of variation. Whereas both designs yield close to nominal coverage, the 1-2-3-3-1 design yields average interval widths that are considerably narrower than the 2-2-2-2-2 design.

Figure 1 suggests that for the CPI data, the variances may be adequately modeled as related to the means by a constant coefficient of variation across all six dose-sex groups. Over all six groups, the average of the ratios $s_{jk}/\bar{u}_{jk}$ is 0.36. Table III displays estimated confidence intervals for the CPI data using the Bailer-model method with $0.36\bar{u}_{jk}$ replacing s_{jk} in (4)-(5). For single AUC's, those intervals are narrower than the ones obtained from the Bailer-Satterthwaite method in five of six cases. None of the intervals from the Bailer-model method include zero. For the differences between groups, the Bailer-Satterthwaite method yields narrower intervals in two of three cases.

Is the Bailer-model method valid? Table V includes the results of a Monte Carlo study to investigate. Data were simulated as described above, according to the 2-2-2-2-2 design. However, confidence intervals were constructed by the Bailer-model method where a constant coefficient of variation was estimated by the average ratio of sample standard deviation to sample mean. Coverages are close to nominal. Average widths for the single-AUC inferences are comparable to the narrow widths of Bailer-Satterthwaite intervals from the 1-2-3-3-1 design. For differences between groups, the model-based widths are slightly larger than those of the Bailer-Satterthwaite intervals with the 1-2-3-3-1 design but still less than those of the Bailer-Satterthwaite intervals with the 2-2-2-2-2 design. The lack of uniform superiority of the Bailer-model method with the real data from a 2-2-2-2-2 design reminds us that confidence intervals are indeed random quantities; and that whereas the Bailer-model method yields narrower intervals on average, for any particular data set the Bailer-Satterthwaite method may produce narrower intervals for some of the groups.

Table V. Comparison of 2-2-2-2-2 and 1-2-3-3-1 Designs, with Bailer-Satterthwaite Method, and Bailer-Model Method for CPI

	Coverage %			Average Width		
	Bailer-Satterthwaite		Bailer-Model	Bailer-Satterthwaite		Bailer-Model
Group	2-2-2-2-2	1-2-3-3-1	2-2-2-2-2	2-2-2-2-2	1-2-3-3-1	2-2-2-2-2
LOW F	96.8	96.5	96.3	3550	1283	1218
LOW M	95.9	96.2	95.7	4790	1739	1499
MID F	95.4	95.8	94.2	15221	5567	4904
MID M	96.5	96.1	94.4	21019	8080	7658
HI F	97.1	95.7	95.3	39811	16810	15925
HI M	96.8	96.1	95.6	79017	29374	26932
MID-LOW	97.5	96.8	96.7	261	148	182
HI-LOW	97.2	96.8	96.9	271	154	185
HI-MID	97.3	95.6	97.5	301	178	219

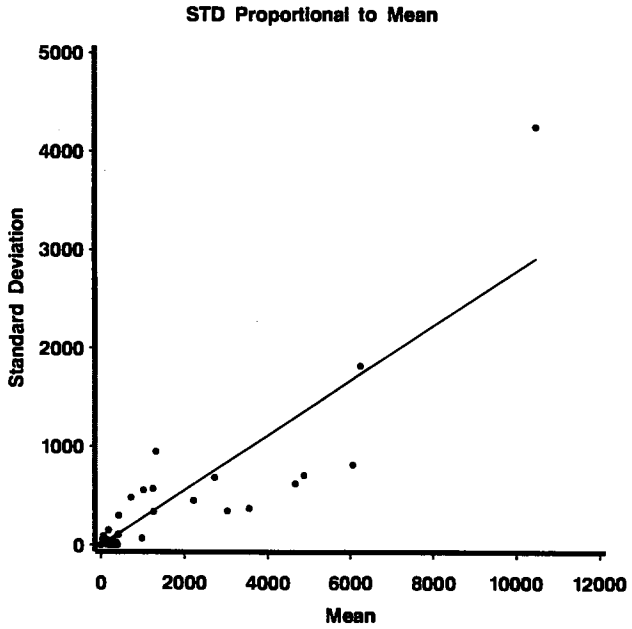

Figure 1: Sample standard deviations vs sample means for all six groups from the CPI study. The line is the no-intercept least-squares regression line.

DISCUSSION

The Satterthwaite approximation of df is easy to compute, and the Bailer-Satterthwaite confidence intervals achieve nominal coverage whereas the intervals from Bailer's method do not. The differences between the two approaches may be small, as shown here with Bailer's and Yuan's data; but, since none of the simple elegance of Bailer's method is sacrificed by performing the extra step, it is reasonable to use the more accurate Bailer-Satterthwaite method. Moreover, with sparse designs, not using the Satterthwaite df can cause serious undercoverage by the confidence intervals.

Our work was motivated by the desire to work with sparse designs, in order to minimize numbers of test animals and assays. With a fixed number of observations to allocate for estimation of AUC's and variances, certain tradeoffs must be accepted. More time points with fewer observations per time point increases the accuracy of $\hat{AUC}$ as an estimate of AUC, but reduces the precision of s^2 ($\hat{AUC}$) as an estimate of σ^2 ($\hat{AUC}$), and vice versa. For our purposes, a five-point estimate of AUC is adequate.

In sparse experimental designs with few replicates and markedly heterogeneous variances and trapezoidal weights, the accurate confidence intervals of the Bailer-Satterthwaite method may be unpleasantly wide. Treatment comparisons may therefore have low power. Wide intervals for single treatments may be of little utility for relating toxicological exposures to human exposures. And if there are positive concentrations among control animals, a confidence interval for an active treatment that includes zero casts doubt on whether the animals in that active group were even exposed. On the other hand, the widths honestly reflect the researcher's uncertainty in the AUC's.

When sufficient prior information is available, the experiment should be designed to increase precision in the estimation of variances,by allocating replicates where variances and trapezoidal weights are large. Or prior knowledge of a variance-mean relationship may make modeling of that relationship feasible. Without that prior knowledge, post hoc examination of the data may suggest a variance-mean model. Of course, the Bailer-model method is not strictly valid if one chooses the variance model post hoc. Nonetheless, data-driven approaches, with inferences drawn from models selected by examining the data, are common in practice. The exploratory nature of the results needs to be recognized. If more confirmatory results are required, follow-up studies can be more efficiently designed, as discussed above, using what has been learned about the variance.

When using the Bailer-model method, care must be taken in fitting the variance model. Although we achieved nominal coverage by estimating the constant coefficient of variation as the average ratio of sample standard deviation to sample mean, we found that estimating it by a no-intercept regression of sample standard deviations on sample means led to under-coverage. The estimation of variation models is an area of active research (5). We restricted attention to approaches that may be less than theoretically optimal in order to retain the simplicity of Bailer's method. More research on variance modeling in sparse-samples for application in toxicokinetics is warranted.

ACKNOWLEDGMENTS

The authors thank Dr. David Cavanagh for directing the toxicity study and Dr. Craig Abolin for the analysis of blood samples. They also thank Dr. William Sallas for useful conversations.

REFERENCES

1. A.J. Bailer. Testing for the equality of area under the curves when using destructive measurement techniques. J. Pharmacokin. Biopharm. 16:303-309 (1988).
2. J. Yuan. Estimation of variance for AUC in animal studies [letter]. J. Pharm. Sci. 82:761-763 (1993).
3. J.R. Nedelman, E. Gibiansky, F.L.S. Tse, and C. Babiuk. Assessing drug exposure in rodent toxicity studies without satellite animals. J. Pharmacokin. Biopharm. 21:323-334 (1994).
4. F.E. Satterthwaite. An approximate distribution of estimates of variance components. Biometrics Bulletin 2:110-114 (1946)
5. M. Davidian and R.J. Carroll. Variance function estimation. *J. Amer. Stat. Assn. 82*:1079-1091 (1987).