

Much Ado about Almost Nothing: Methods for Dealing with Limited Data

Stephen W. Looney¹ and Courtney E. McCracken²

1. Department of Population Health Sciences, Augusta University, 1120 15th Street, AE-1014, Augusta, GA 30912-4900, USA

2. Department of Pediatrics, Emory University, 1760 Haygood Drive NE, Suite W454, Atlanta, GA 30322, USA

Abstract: Applied statisticians are often confronted with statistical inference problems dealing with situations in which there appear to be no data, or data of only limited usefulness. For example, when attempting to find a confidence interval for a binomial proportion, the sample may contain no successes. Such a scenario could be encountered when attempting to estimate the incidence of an extremely rare side effect associated with the administration of a newly developed drug. In this article, we use examples for our experiences working with scientific investigators and describe several scenarios in which there appeared to be no useful data, or data of only limited usefulness. We describe the methods we prefer for analyzing the data in these situations and illustrate their application using the actual data from the investigations we participated in.

Key words: Binomial proportion, correlation analysis, exact methods, kappa coefficient, LOD (limit of detection), OR (odds ratio), RR (risk ratio), *t*-test.

1. Introduction

First of all, we must apologize to Dennis Helsel for “stealing” part of the title of our article from him. He is the author of *Nondetects and Data Analysis* [1], which we consider to be the definitive text on the analysis of data that include observations below the LOD (limit of detection) of a measuring device. (We will refer to such observations as “nondetects”, or ND’s). In 2010, he published a commentary entitled “Much Ado about Next to Nothing: Incorporating Nondetects in Science” [2], which addresses many of the issues that we discuss in this article. (Of course, we also owe a debt of gratitude to the Bard of Avon for our title).

In this article, we address an issue that applied statisticians are often confronted with; namely, statistical inference problems in which there appear to be no useful data, or data of only limited usefulness. We illustrate each of the following scenarios with real

data and describe the methods that we prefer for analyzing the data in these situations: (1) Finding a C.I. (confidence interval) for a binomial proportion when the sample contains no successes (or failures); (2) Finding a C.I. for a RR (risk ratio) when one of the groups experiences none of the outcome of interest; (3) Finding a C.I. for an OR (odds ratio) when one of the cells in the 2×2 table is empty; (4) Estimating the OR when there are incomplete pairs in a matched case-control study; (5) Combining data from paired and unpaired *t*-tests in the same analysis; (6) Measuring agreement with Cohen’s kappa where there is an empty cell or empty row in the 2×2 table; and (7) Incorporating ND’s into a correlation analysis.

2. Estimating a Binomial Proportion when There Are No Successes

This situation could occur, for example, when we wish to estimate the incidence of an extremely rare side effect associated with the administration of a drug or use of a device. For example, one of the authors (SL) was involved in a clinical study in which a dentist wanted to examine the effects of electromagnetic

Corresponding author: Stephen W. Looney, Ph.D, research fields: analysis of messy data, analysis of biomarker data, multivariate data analysis.

interference on a neurostimulator during the operation of three dental devices (electric pulp tester, apex locator, electrocautery unit) [3]. After $n = 70$ independent trials, there were no failures of the neurostimulators. The researcher wanted to know: “What is a reasonable upper bound for the probability of failure of the neurostimulator?” If one formulates this in the context of the binomial distribution, letting $\pi =$ the probability of failure, then the problem becomes one of finding the upper limit of a 1-sided 95% C.I. for a binomial proportion: $(0, \pi_u)$. The most commonly used method, the Wald interval, yields uninformative results:

$$\hat{\pi} = \frac{0}{70} = 0$$

$$ASE(\hat{\pi}) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{70}} = \sqrt{\frac{0(1-0)}{70}} = 0, \pi_u = 0$$

(In this formula and throughout the remainder of this article, “ASE” will be used to indicate “approximate standard error”).

We evaluated 8 methods that can be used to find confidence limits for a binomial proportion under the assumption that $x =$ the number of successes in a binomial sample of size n that has already been observed to be zero [4]: (1) The mid-P interval [5]; (2) The Wilson (score) interval [6]; (3) The SAIFS (single

augmentation with an imaginary failure or success) interval [7]; (4) The Bayes-Laplace HPD interval [8]; (5) The Clopper-Pearson (C-P) exact interval [9]; (6) The Poisson interval, also known as the rule of 3 for 95% confidence intervals [10]; (7) The continuity-corrected Wilson interval [11]; and (8) The Agresti-Coull interval [12]. Table 1 contains the upper 95% confidence limit based on the observed sample of $n = 70, x = 0$ for each of these methods.

Methods typically used to evaluate C.I. procedures such as expected interval length and coverage probability do not apply in the present case because it is assumed that the data have already been observed; namely, that $x = 0$. Alternative methods for evaluating C.I. procedures that do apply in this situation include p -confidence and p -bias [13] and observed interval length.

Compared to the other methods we evaluated, the C.I.’s based on the Clopper-Pearson exact method is preferred when $x = 0$ for the following reasons: (1) They are moderate in length; (2) The required sample size is comparable to other methods; (3) They are easy to compute; (4) There are closed-form expressions for π_u and the sample size required to yield one-sided C.I. of any length; (5) They are equivalent to a Bayesian prediction interval based on the Jeffreys prior;

Table 1 Comparison of methods for finding the upper 100 (1- α)% confidence limit, π_u , for a binomial proportion when $x = 0, n = 70, \alpha = 0.05$.

Method	Formula for upper limit	π_u
Mid-P [5]	$1 - (2\alpha)^{1/n}$	0.032
Wilson (score) [6]	$Z_{1-\alpha}^2 / (n + Z_{1-\alpha}^2)$	0.037
SAIFS [7]	$(1 + t_{n-1, 1-\alpha}) / (n+1)$	0.038
Bayes-Laplace [8]	$1 - (\alpha)^{1/(n+1)}$	0.041
C-P Exact [9]	$1 - (\alpha)^{1/n}$	0.042
Rule of 3 [10]	$[-\ln(\alpha)] / n$	0.043
Continuity-corrected Wilson [11]	$\{[0.5 + (Z_{1-\alpha}^2 / 2) + Z_{1-\alpha}] \sqrt{0.5 - (0.25 / n) + (Z_{1-\alpha}^2 / 4)}\} / (n + Z_{1-\alpha}^2)$	0.050
Agresti-Coull [12]	$[2 / (n + 4)] + Z_{1-\alpha} \sqrt{[2 / (n + 4)][1 - 2 / (n + 4)] / (n + 4)}$	0.058

and (6) They always have p -confidence equal to the nominal confidence coefficient. See Ref. [4] for full details.

3. Estimating a RR when There Are No Events in One of the Exposure Groups

In a study of a behavioral intervention for depression in nursing homes [14], the following results were obtained when considering falls as an adverse event:

The goal of this analysis was to estimate the association between treatment group (intervention vs. control) and outcome (fall vs. no fall) using the RR and the associated 95% C.I. For the data in Table 2, using the usual “ a, b, c, d ” notation for a 2×2 table, the standard approach [5] yields:

$$\widehat{RR} = \frac{a / (a + b)}{c / (c + d)} = \frac{5 / 29}{0 / 24} = \infty, \quad (1)$$

$$ASE[\log(\widehat{RR})] = \sqrt{\frac{1}{a} + \frac{1}{c}} = \sqrt{\frac{1}{5} + \frac{1}{0}} = \infty.$$

An approximate 95% C.I. for the log RR is given by $\log(\widehat{RR}) \pm z_{\alpha/2} ASE[\log(\widehat{RR})]$ where $z_{\alpha/2}$ denotes the upper $\alpha/2$ percentage point of the standard normal distribution. The endpoints of this C.I. are then exponentiated to obtain an approximate 95% C.I. for

the RR. This yields a 95% C.I. (RR) of $(-\infty, \infty)$ for the data in Table 2. The method of Agresti and Caffo [15] that can be used for statistical inference for the risk difference when zero events are observed in one of the groups that can be adapted to perform inference for the RR. This method is based on applying the formulas in Eq. (1) to the data in the following “amended” table, obtained by adding 1 to each cell in Table 2.

Using the data in Table 3, we obtained $\widehat{RR} = 5.3$, 95% C.I. (0.65, 39.16). Exact methods are also available for estimating the true RR when there are no events in one of the two groups [16] and these have been implemented in StatXact. Using StatXact, the following results were obtained: $\widehat{RR} = \infty$, exact 95% C.I. (1.14, ∞). Carter et al. [17] proposed an alternative method for estimating the RR that is based on using the ratio of the median unbiased estimates of the proportions in each treatment group to obtain a point estimate of the RR and the deterministic bootstrap to find an approximate C.I. (RR). For the data in Table 2, $\widehat{RR} = 12.38$, 95% C.I. (1.52, 23.44). Note that the conclusions differ for the 3 intervals in terms of the test of $H_0: RR = 1$, where RR denotes the true value of the RR (Table 4).

Table 2 Association between risk of fall and intervention for depression ($n = 53$).

Treatment group	One or more falls	No falls
Intervention	5 (a)	24 (b)
Control	0 (c)	24 (d)

Table 3 Amended table corresponding to Table 2.

Treatment group	One or more falls	No falls
Intervention	6	25
Control	1	25

Table 4 Comparison of methods for finding a confidence interval for the RR when there are no events in one of the groups.

Method	\widehat{RR}	95% C.I.	Conclusion
Amended	5.3	(0.65 – 39.16)	Fail to reject
Exact	∞	(1.14, ∞)	Reject
Carter et al. (2010)	12.38	(1.52, 23.44)	Reject

4. Estimating the OR for a 2 × 2 Table with an Empty Cell

In a study of the association between positive toxicology screens and clinical outcome among patients hospitalized with traumatic injuries [18], the following 2 × 2 table was obtained.

Using the usual “*a, b, c, d*” notation as in Table 5, the OR is typically estimated as follows [5]:

$$\widehat{OR} = \frac{ad}{bc},$$

$$ASE[\log(\widehat{OR})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}. \quad (2)$$

An approximate 95% C.I. for the log OR is given by $\log(\widehat{OR}) \pm z_{\alpha/2} ASE[\log(\widehat{OR})]$. The endpoints of this C.I. are then exponentiated to obtain an approximate 95% C.I. for the OR. If any of the cells in the 2 × 2 table are 0, then either $\widehat{OR} = 0$ or $\widehat{OR} = \infty$, depending on whether the 0 cell occurs in the main diagonal or the off-diagonal, respectively. Furthermore, $ASE[\log(\widehat{OR})] = \infty$ if any of the cells in the 2 × 2 table are zero.

For a 2 × 2 table with a zero in one of the cells, Agresti [5] recommends the use of the “slightly amended” estimator of the OR: $\widehat{OR} = \frac{(a+.5)(d+.5)}{(b+.5)(c+.5)}$. The formulas in Eq. (2) are then applied after adding .5 to each cell count.

For the data in Table 5, $\widehat{OR} = 0$ and $ASE[\log(\widehat{OR})] = \infty$. The “Agresti method” is based on the following amended table:

Using the data in Table 6, we obtain $\widehat{OR} = 0.11$, 95% C.I. (0.01, 1.81). Exact methods are also available for estimating the true OR when there are empty cells [19] and these have been implemented in StatXact. For the data in Table 5, StatXact yields $\widehat{OR} = 0$, exact 95% C.I. (0.00, 0.72). Note that the two approaches differ in terms of the conclusion of the test of $H_0: OR = 1$, where *OR* denotes the true value of the OR.

5. Estimating the OR from Incomplete Pairs in a Matched Case-Control Study

A commonly used study design for examining the association between occurrence of disease and exposure to a risk factor that is the matched case-control study. In epidemiological research, it is usually desirable to control for various confounders; one approach for doing this is to match each case (a study participant with the disease) to a control (a participant without the disease) in terms of age, race, sex, socio-economic status, etc. However, difficulties in estimating the true OR and performing inference for it occur when the exposure status is unknown for one of the individuals in one or more matched pairs. When this occurs, we say that the matched pair is “incomplete”. Case-control studies in which incompletely matched cases and controls were present include London et al. [20] and Pike et al. [21].

In the presence of incomplete matched pairs as described above, authors have either removed the incomplete pairs from the analysis [21] or ignored the

Table 5 2 × 2 table showing association between positive screen for cocaine and risk of death following traumatic injury.

	Outcome	
	Death	No death
Cocaine Positive	0 (a)	110 (b)
Cocaine Negative	30 (c)	739 (d)

Table 6 Amended table corresponding to Table 5.

	Outcome	
	Death	No death
Cocaine positive	0.5	110.5

Cocaine negative

30.5

739.5

matching and treated all cases and controls as if they were independent [20]. Neither of these approaches is optimal; in the former, part of the sample has been deleted, thereby reducing the sample size and decreasing the precision of the OR estimate, which leads to confidence intervals that are too wide and hypothesis tests that have reduced power. In addition, bias may be introduced by deleting part of the sample. In the latter approach, treating the paired cases and controls as if they were independent results in distorted variance estimates. This can lead to invalid results when performing statistical inference for the OR.

“Matching” may also result from the fact that the same study participants were observed or measured on two different occasions. In an unpublished British study, 5,400 asthma patients were prescribed either a generic or proprietary version of an asthma inhaler. Patients were categorized by age (5-15 years or 16-64 years) at time of prescription. The purpose of the study was to estimate the association between age group and type of inhaler. Hypothetical data similar to those obtained in the research study are

provided in Table 7.

The problem is that 1,300 patients participated in the study when they were in the 5-15 year age group and also when they were in the 16-64 year age group, i.e., these 1,300 were “matched” (the remaining 2,800 participated in the study when they were in only one of the age groups).

We believe that the first step in the proper analysis of a mixed sample of matched and unmatched cases and controls is to divide the sample into the paired subsample and the unpaired subsample. The paired subsample is presented in a 2×2 table showing the results for each matched pair. For example, in the asthma study, there were 100 participants who used the generic inhaler when they were in the 5-15 age group and also when they were in the 16-64 age group (Table 8). Using standard notation for the cell counts in the 2×2 table for matched pairs, the OR was estimated by $\frac{b_0}{c_0} = \frac{500}{400} = 1.250$. Thus, in the paired subsample, the odds of using a proprietary inhaler are increased by a factor of 1.25 in the 16-64 age group relative to the 5-15 age group.

Table 7 Hypothetical data on association between age group and type of asthma inhaler prescribed.

Age group	Type of asthma inhaler	
	Generic	Proprietary
5-15 years	900	1,300
16-64 years	1,400	1,800

Table 8 Hypothetical data-paired cases and controls only.

		Aged 16-64	
		Generic	Proprietary
Aged 5-15	Generic	100 (a ₀)	500 (b ₀)
	Proprietary	400 (c ₀)	300 (d ₀)

Table 9 Hypothetical data-independent cases and controls only.

Age group	Type of asthma inhaler	
	Generic	Proprietary
5-15 years	300 (a ₁)	600 (b ₁)

16-64 years	900 (c ₁)	1,000 (d ₁)
-------------	--------------------------	----------------------------

In the unpaired sample, the cases and controls are assumed to be independent. In the asthma study, there were 900 participants who were in the study when they were in the 5-15 age group, but were not in the study when they were in the 16-64 age group. Similarly, there were 1,900 participants who were in the study when they were in the 16-64 age group, but not when they were in the 5-15 age group (Table 9). The 900 participants were assumed to be independent of the 1,900 and therefore the OR was estimated using

$$\widehat{OR}_2 = \frac{a_1 d_1}{b_1 c_1} = \frac{(300)(1000)}{(600)(900)} = 0.56.$$

the standard method:

Thus, in the unpaired subsample, the odds of using a proprietary inhaler were decreased by a factor of 0.56 in the 16-64 age group relative to the 5-15 age group.

Several authors have proposed methods that make use of all the data when estimating the OR in matched case-control studies with incomplete pairs. For example, Haber and Chen [22] proposed “matched” and “unmatched” estimators and Huberman and Langholz [23] proposed the MI (missing-indicator) method. We proposed an alternative estimator and compared it with Haber and Chen’s estimators and the MI method [24].

Our method is computationally simpler than any of the other proposed methods and is based on a weighted average of the estimated OR’s calculated using the paired and unpaired subsamples as described above. It was motivated by the well-known result that the optimal weights for combining two unbiased estimators of an unknown parameter are proportional

to the reciprocals of the respective variances of the two estimators [25].

Using the notation for the cell counts in the paired subsample given in Table 8, the usual estimate of the log OR based on the completely matched pairs, is

given by $\widehat{\beta}_1 = \log(OR_1) = \log\left(\frac{b_0}{c_0}\right)$ with estimated variance $V_1 = \widehat{Var}(\widehat{\beta}_1) = \frac{1}{b_0} + \frac{1}{c_0}$.

Using the notation for the cell counts for the unpaired subsample in Table 9, the usual estimate of the log OR is given by

$$\widehat{\beta}_2 = \log(OR_2) = \log\left(\frac{a_1 d_1}{b_1 c_1}\right)$$

with estimated variance $V_2 = \widehat{Var}(\widehat{\beta}_2) = \frac{1}{a_1} + \frac{1}{b_1} + \frac{1}{c_1} + \frac{1}{d_1}$.

The log OR estimator proposed in Ref. [24] is given by $\widehat{\beta} = w_1 \widehat{\beta}_1 + w_2 \widehat{\beta}_2$ with $w_i = V_i^{-1} / (V_1^{-1} + V_2^{-1})$, where, $V_i = \widehat{Var}(\widehat{\beta}_i)$, $i = 1, 2$.

A straightforward derivation shows that $\widehat{Var}(\widehat{\beta}) = V_1 V_2 / (V_1 + V_2)$. A Wald-type 100(1- α)% C.I. for the true log OR is then given by:

$$\widehat{\beta} \pm z_{\alpha/2} \sqrt{\widehat{Var}(\widehat{\beta})}. \quad (3)$$

An approximate 100(1- α)% C.I. for the true OR is obtained by exponentiating the endpoints in Eq. (3). The results of using each of the various estimators described here with the asthma inhaler data are summarized in Table 10.

The OR estimate based on ignoring the matching and the OR estimate based on our method are similar in magnitude and the C.I. is approximately the same width; however, the “matching ignored” method

Table 10 Summary of results for estimating the OR using incompletely paired data.

Method	Sample size	\widehat{OR}	95% C.I.
Matched pairs only	1,300	1.25	(1.10, 1.43)
Unmatched cases and controls only	2,800	0.56	(0.47, 0.66)
Combined (matching ignored)	5,400	0.89	(0.80, 0.99)
Matched estimator [22]	5,400	0.88	(0.81, 0.95)

Unmatched estimator [22]	5,400	0.89	(0.79, 1.00)
Missing indicator method [23]	5,400	0.909	(0.821, 1.006)
New method [24]	5,400	0.913	(0.824, 1.012)

indicates statistical significance, whereas our method does not. The results for our method and the MI method are almost identical. Haber and Chen’s “unmatched” estimator yields similar results, except that the estimated O.R. is slightly smaller and the C.I. is about 15% wider than for our method. Our estimator, the MI estimator, and Haber and Chen’s unmatched estimator are the only approaches that do not indicate a statistically significant association between age group and type of inhaler. As can easily be seen from the results in this table, different analysis methods can yield very different estimates of the O.R. and very different conclusions when the data consist of a mixed sample of paired and unpaired cases and controls.

In an extensive simulation study [24], we evaluated the bias, mean squared error, CIW (C.I. width), and CP (coverage probability) of both of Haber and Chen’s methods, the MI method, and our method. For most of the combinations of simulation parameters that we considered, our estimator performed at least as well as each of the other estimators. In particular, our estimator generally had smaller bias and smaller mean squared error, the CP of the C.I.’s based on our estimator was generally closer to the nominal level and the CIW for our C.I.’s tended to be smaller when compared with the other estimators. See Ref. [24] for full details.

6. The Two-Sample Problem with Mixed Samples of Correlated and Uncorrelated Data

In a study of the quality of life of lung cancer patients in the last few days of life [26], Karnofsky Ratings (a measure of functional status) were obtained on 37 patients. Out of the 37, 9 were observed on both days, 23 were observed only on their last day of life,

and 28 were observed only on the day before they died. Let “Group 1” denote those patients who were observed on the day before they died, and let “Group 2” denote those patients who were observed on the day they died. The purpose of the study was to determine if mean functional status was significantly different between Groups 1 and 2.

This is an example of a two-sample problem in which the data consisted of one subsample in which the study participants in Group 1 were independent of those in Group 2, and another subsample in which the participants in Group 1 was paired with those in Group 2 (and hence were not independent). Other studies in which the study sample consisted of subsamples of correlated and uncorrelated data include Dimery et al. [27], Nurnberger et al. [28] and Steere et al. [29]. In each of these studies, the fact that some of the observations were paired between the two treatment groups was ignored and the samples from the two groups were analyzed as if they were independent. Thus, the correlation between the paired observations was not incorporated into the analysis. Ignoring this dependence between some of the observations in the study sample can lead to distorted estimates of the variance of the difference between the sample means in the two groups, which can seriously affect statistical inference for the difference in population means [30].

The most commonly used method for analyzing a “combined” sample of this type is to use the unpaired t -test to find a C.I. or test hypotheses about the true difference in population means. This approach is based on the assumption that there is no correlation between the observations in Groups 1 and 2, and we will refer to it as the “uncorrected t -test”. Another approach is to ignore the paired observations and use the unpaired t -test after removing the correlated data,

and we will refer to it as the “unpaired t -test”. A third method that could be used would be to ignore the Groups 1 and 2 observations that are independent of each other and perform the paired t -test after removing the uncorrelated data; we will refer to this as the “paired t -test”. All of these approaches are suboptimal. Statistical inference based on the uncorrected t -test leads to distorted estimates of the variance of the difference between the sample means in the two groups; using either the unpaired t -test or the paired t -test would involve removing part of the sample, which leads to decreased precision of the sample difference in group means, resulting in wider C.I.’s and lower power of statistical tests for the true difference in group means.

We proposed an alternative method, “the corrected z -test” for analyzing combined samples of correlated and uncorrelated observations in the two-sample problem [31]. It makes use of all the data and takes into account the correlation between the paired observations. In a limited simulation study using only normally distributed data, we found that the corrected z -test generally performed well in terms of Type I error and power for sample sizes of 50 in each of the two groups.

Let n_1 = number of unpaired Group 1 observations, M_1 = mean of all $(n_1 + n)$ Group 1 observations, and S_1^2 = variance of all $(n_1 + n)$ Group 1 observations. Define similar notation for the observations in Group 2: n_2 , M_2 , and S_2^2 . Let n = number of paired observations, and let $S_{u,v}$ = covariance of the n paired observations. We proposed the “corrected z -test”, where the test statistic is given by:

$$Z_{corr} = \frac{M_1 - M_2}{\sqrt{\text{Var}(M_1 - M_2)}}$$

and

$$\text{Var}(M_1 - M_2) = \frac{S_1^2}{n_1 + n} + \frac{S_2^2}{n_2 + n} - \frac{2nS_{uv}}{(n_1 + n)(n_2 + n)}. \quad (4)$$

By the Central Limit Theorem and consistency, Z_{corr} has a standard normal limiting distribution

under the null hypothesis that the difference in population means is zero [31].

As pointed out by Uddin and Hasan [32], it is possible under extreme conditions (e.g., large n relative to n_1 and n_2 and large S_{uv}), the variance estimate in Eq. (4) can be negative. They proposed a modification to the estimated variance in Eq. (4), resulting in their test statistic z_c .

In their simulation study, Uddin and Hasan found that the z_c test performed well in terms of Type I error and power relative to the other estimators they considered.

The corrected z test in its original form was compared to several other methods by Guo and Yuan [33]. They found that the corrected z test generally performed well compared to the other methods for moderate sample sizes ($n_1 \geq 10$, $n_2 \geq 10$, and $n = 20$). However, they found that the test based on a modified maximum likelihood estimator (MMLE) was generally preferable for moderate n when the data were normally distributed. When the data were not normally distributed, they recommended their “optimal pooled t -test”. When the sample size was small, they recommended their optimal pooled t -test when there were missing data at both time points and their “paired t -test” when there were missing data at only one time point. A SAS program for performing each of these procedures is available from the first author. In the Karnofsky example described above, the corrected z -test [31] yielded $z_{corr} = 2.50$, $p = 0.012$; the MMLE [33] yielded $t_{MMLE} = 2.93$, $d.f. = 8$, $p = 0.019$; and the modified corrected z -test [32] yielded $z_c = 2.46$, $p = 0.014$.

7. Measuring Agreement when There Is an Empty Cell or Row in the 2×2 Table

We are often interested in measuring agreement between two dichotomous variables. Examples include inter- and intra-rater reliability in which we wish to measure the extent of agreement between two raters (examiners) or between two occasions with the

same rater, and so-called method comparison studies in which we wish to measure the extent of agreement between two measurement methods, neither of which is the gold standard.

In an unpublished study, the investigators wished to examine the agreement between two methods of diagnosing abnormal BMD (bone mineral density) in children with asthma: (1) The 1st percentile method; and (2) The *z*-score method. The gold standard method of measuring BMD was not available. The agreement data between the two methods when only the male children with asthma are considered are presented in Table 11.

One way to measure agreement between the two methods using the data in Table 11 is to use the total percentage of agreement, also known as the ICA (Index of Crude Agreement). Using the standard “*a, b, c, d*” notation as in Table 11, the ICA is given by

$$p_0 = \frac{a + d}{n}.$$

Another measure of agreement between two dichotomous methods of classification is Cohen’s kappa, given by $\kappa = \frac{p_0 - p_e}{1 - p_e}$, where p_e = the estimated percentage agreement between the two methods that “can be attributed to chance” [34]. Using the same notation as in Table 11,

$$p_e = \frac{(a + b)(a + c) + (b + d)(c + d)}{n^2}.$$

It can be shown that $\kappa = 1$ if there is perfect agreement between the two methods of classification (i.e., $b = c = 0$), $\kappa = -1$ if there is perfect disagreement between the two methods ($a = d = 0$), and $\kappa = 0$ if $p_0 = p_e$. Landis and Koch [35] provided guidelines for interpreting the magnitude of κ : if $\kappa < 0.00$, this is interpreted as “poor” agreement; values between 0.00 and 0.20 indicate “slight” agreement; values between 0.21 and 0.40 indicate “fair” agreement; 0.41-0.60 indicates “moderate” agreement; 0.61-0.80 indicates “substantial” agreement; and values between 0.81 and 1.00 indicate “almost perfect” agreement. For the agreement data presented in Table 11, $p_0 = 38/44 = 0.86$, $p_e = 0.70$, and $\kappa = 0.54$, indicating moderate agreement according to the Landis-Koch criteria.

Cohen’s kappa is the most commonly used method of agreement between two dichotomous classifiers; however, there are problems with kappa when there are limited agreement data: (1) If either method classifies no subjects into one of the two categories, $\kappa = 0$; (2) If there are no agreements for one of the two categories, $\kappa < 0$.

To illustrate problem Eq. (1), consider Table 12, which contains the agreement data between the two methods of diagnosing abnormal BMD when only female children with asthma are considered.

Table 11 Agreement between the 1st percentile and Z-score methods of diagnosing abnormal bone mineral density in male children with asthma.

Z-score method	1st percentile method	
	Positive	Negative
Positive	5 (a)	4 (b)
Negative	2 (c)	33 (d)

Table 12 Agreement between two methods of diagnosing abnormal bone mineral density in female children with asthma.

Z-score method	1st percentile method	
	Positive	Negative
Positive	0	0
Negative	3	25

Table 13 Agreement between two methods of diagnosing abnormal bone mineral density for children on low-dose

corticosteroid therapy.

Z-score method	1st percentile method	
	Positive	Negative
Positive	0	2
Negative	4	26

For these data, $p_0 = 25/28 = 0.89$, $p_e = 0.89$, and $\kappa = 0.00$, which indicate poor agreement using the Landis-Koch criteria.

To illustrate problem Eq. (2), consider Table 13, which contains the agreement data between the two methods when only those children on low-dose corticosteroid therapy are considered.

For these data, $p_0 = 26/32 = 0.81$, $p_e = 0.83$, and $\kappa = -0.09$, which also indicate poor agreement using the Landis-Koch criteria.

Several alternatives to kappa have been proposed for measuring agreement between two dichotomous classifiers. One that we have found to be particularly useful is the prevalence-adjusted and bias-adjusted kappa (PABAK) statistic [36]. Calculation of PABAK is straightforward and involves replacing a and d in the original 2×2 table by their average, $r = \frac{a+d}{2}$ and replacing b and c by their average, $s = \frac{b+c}{2}$. The PABAK statistic is then given by the value of kappa for this “new” table, where the “new” $p_0 = 2r/n$ and the “new” $p_e = 0.5$:

$$PABAK = \frac{(2r/n) - 0.5}{1 - 0.5} = 2p_0 - 1.$$

PABAK can also be shown to be equal to proportion of “agreements” minus the proportion of “disagreements”:

$$PABAK = \frac{(a+d)}{n} - \frac{(b+c)}{n}.$$

The value of PABAK is interpreted using the Landis-Koch criteria for interpreting the value of κ .

For the data in Table 12, $PABAK = 2(0.89) - 1 = 0.78$, which indicates “substantial agreement” between the two methods; for the data in Table 13, $PABAK = 2(0.81) - 1 = 0.62$, which also indicates “substantial

agreement”. To us, these PABAK values provide a much more accurate representation of the degree of agreement between the two methods of diagnosing low BMD in children than the corresponding kappa values (0.00 and -0.09, respectively).

8. Estimating Correlation in the Presence of Non-detects

Looney and Hagan [37] considered the correlation between hippuric acid (Y) and *ortho*-Cresol (X) concentrations in urine samples of 54 individuals exposed to toluene using data taken from Amorim and Alvarez-Leite [38]. Both X and Y were subject to non-detects (ND’s). Let LOD_x denote the LOD of X and let LOD_y denote the LOD of Y . Out of the $n = 54$ pairs of (X, Y) observations, there were 15 subjects with $X \geq LOD_x$ and $Y \geq LOD_y$, 35 subjects with $X < LOD_x$ and $Y \geq LOD_y$, and 4 subjects with $X < LOD_x$ and $Y < LOD_y$. Thus, there were only 15 subjects in the “complete sample”, which is, those for whom their X and Y values were both no smaller than the respective LODs for X and Y .

Perhaps the most commonly used method of dealing with ND’s in correlation analysis is to remove the observations with NDs in either X or Y and base the statistical analysis on the complete sample. Lagorio et al. [39] used this approach in their evaluation of trans, trans muconic acid (t, t -MA) as a measurement for low-level benzene exposure. Another commonly used approach is simple substitution; in other words, replace the observations below the LOD with some substitute value and then apply the desired statistical analysis to the new sample of data that contains the substituted values in place of the NDs. The most commonly substituted values include the LOD [38, 40] and $LOD/2$ [41].

Another method that has been proposed for dealing with NDs is the nonparametric approach, in which one treats all NDs as if they were tied at some value just below the LOD. Thus, if one wished to correlate two measurements X and Y , at least one of which was undetectable in some samples, one could calculate Spearman's r_s using the ranks of the X and Y values based on the entire data set, where all NDs for each measurement have been assigned the smallest mid-rank. Kendall's coefficient has also been proposed for use when correlating measurements subject to LOD [1].

In most statistical analyses of data subject to a LOD, the ND's are treated as left-censored at the LOD. If both X and Y are subject to left-censoring and the vector (X, Y) can be assumed to have a bivariate (BVN) distribution, the maximum likelihood (ML) approach proposed by Lyles et al. [42] can be used. It is important to note that ND's cannot be assumed to be missing at random.

In her PhD dissertation research, the 2nd author performed a comprehensive comparison of eight methods for dealing with ND's when estimating a population correlation, ρ [43]. The methods she compared included (1) The ML approach [42]; (2) Simple substitution with several different values, including LOD , $LOD/2$, and $LOD/\sqrt{2}$; (3) Complex substitution [44], in which each ND among the x -values is replaced by $E(X_i | X_i < LOD_x)$ and each ND among the y -values is replaced by $E(Y_i | Y_i < LOD_y)$; (4) random imputation from a uniform distribution, in which each ND among the x -values is replaced by a randomly selected value from the interval $[0, LOD_x]$, and each ND among the y -values is replaced by a randomly selected value from the interval $[0, LOD_y]$; and (5) Spearman's correlation. For estimation methods (1)-(4) above, a 2nd-order Fisher z -transformation, which provides a more accurate estimate of the variance of $z(\hat{\rho})$, was used to find a 95% C.I. for the true value of the correlation between X and Y . Confidence intervals

based on this method have improved coverage probabilities relative to those based on the usual Fisher z -transformation, and the improved transformation poses no computational difficulties. Details of this method are provided by Li et al. [45]. For estimation method (5), both the jackknife and approximate bootstrap confidence interval (ABC) methods were evaluated. All of these estimation methods were compared under conditions of large censoring proportions, unbalanced censoring proportions, small to moderate n , and non-BVN data.

An extensive Monte Carlo simulation study was undertaken by McCracken to compare the various point estimates and corresponding 95% C.I. procedures. Detailed results from this Monte Carlo study can be found in Ref. [43] and a summary is provided in Section 4.8 of Ref. [46]. The sample sizes that McCracken considered ranged from $n = 20$ to $n = 500$, the true correlation between X and Y prior to censoring ranged from $\rho = -0.9$ to $\rho = 0.9$, and the censoring proportions on X (p_1) and Y (p_2) ranged from $(p_1, p_2) = (0,0)$ to $(p_1, p_2) = (0.90, 0.90)$. Both balanced and unbalanced combinations of censoring proportions were considered. For example, unbalanced combinations included $(10, 0)$, $(10, 75)$, $(90, 45)$, and $(90, 0)$. The true bivariate distribution of X and Y was also evaluated as a simulation parameter. Data were generated from the BVN (bivariate normal), BVG (bivariate gamma), and BVB (bivariate beta) distributions. The non-BVN distributions were chosen to represent substantial departures from the BVN, as measured by Mardia's coefficients of multivariate skewness and kurtosis, denoted by $\beta_{1,p}$ and $\beta_{2,p}$, respectively. For the bivariate normal, $\beta_{1,2} = 0$ and $\beta_{2,2} = 8$. For the BVG used in the simulation study, $\beta_{1,2} = 3.5$ and $\beta_{2,2} = 12$ and, for the BVB, $\beta_{1,2} = 3$ and $\beta_{2,2} = 10$. For further details on the simulation study, see Ref. [43].

Each of the point estimates and corresponding 95% C.I. procedures described above were evaluated in

terms of the following criteria: (1) Bias (and absolute bias); (2) Median absolute deviation; (3) CIW; and (4) C.I. CP. McCracken’s results showed that when the data were generated from a BVN distribution, the ML method was superior to all other estimation methods under all conditions considered. When the data were generated from the non-BVN distributions, the performance of the ML method was still superior to the other estimation methods except under some scenarios involving moderate to large n , small $|\rho|$ and very light censoring combinations. Overall, the MLE (maximum likelihood estimate) of ρ had the best overall performance in terms of all criteria that were considered even when the assumption of BVN was violated. This was somewhat surprising since the MLE was derived under the assumption that (X, Y) had a BVN distribution. For the Spearman coefficient, the C.I. based on the jackknife was comparable to that based on the ABC method in terms of performance and was computationally simpler, so McCracken recommended that the jackknife could be used to find a 95% C.I. for the Spearman correlation. As expected, point estimates and C.I.’s based on Spearman’s coefficient generally performed as well as the ML method (and sometimes better) with non-BVN data, but were inferior to the MLE for several settings of the simulation parameters when the data were generated from a BVN distribution. All of the substitution methods (LOD, LOD/2, etc.) described previously had inferior overall performance relative to the MLE and Spearman’s coefficient.

Consider once again the study by Amorim and Alvarez-Leite [38], there is strong evidence against the bivariate normality assumption for these data (Shapiro-Wilk $p < 0.0001$ for both the 15 *o*-cresol values and the 50 hippuric acid values). Table 14

contains a summary of the results of various methods that could be used to estimate the correlation between the hippuric acid and *ortho*-Cresol concentrations. As can be seen from the table, the results for the ML approach differ very little from those produced by the substitution methods. However, there is quite a discrepancy between the results obtained using the ML method and those based on either the Spearman or Kendall coefficients. Note that the censoring proportions in this example are $p_1 = 0.07$ (4/54) for hippuric acid and $p_2 = 0.72$ (39/54) for *o*-cresol. For data that do not appear to follow a BVN distribution (as is the case in this example), McCracken’s simulation results indicated that C.I.’s based on either the MLE or Spearman’s r_s maintained acceptable CP under the censoring proportions she considered that were closest to (0.07, 0.72), namely (0.10, 0.70) (median CP over all simulation parameter settings was 93.9% for ML and 92.9% for r_s). Her results indicated that CIs based on the MLE maintained acceptable CP (93.4%) when the true value of ρ was approximately 0.75, which seems to be a reasonable assumption based on the results in Table 14 ($\hat{\rho}_{ML} = 0.79$) Similarly, C.I.’s based on Spearman’s r_s maintained acceptable CP (93.4%) when the true value of ρ_s was approximately 0.5, which appears to be a reasonable assumption based on the results in Table 14 ($r_s = 0.58$). Finally, from the simulation results for non-BVN data, C.I.’s based on either the MLE (93.8%) or Spearman’s r_s (93.7%) maintained acceptable CP when n was approximately 50, which it is for this set of data ($n = 54$). Thus, based on McCracken’s simulation results, we have no reason to doubt the validity of either the ML-based C.I. or the r_s -based C.I. Given the strong departure from normality for these data, we recommend that Spearman’s coefficient be used:

Table 14 Comparison of estimated correlations using the data in Ref. [38].

Method	n	$\hat{\rho}$	95% C.I.	CIW
Substitute zero	54	0.79	(0.67, 0.88)	0.21
Substitute LOD/2	54	0.79	(0.65, 0.87)	0.22
Substitute LOD	54	0.77	(0.63, 0.85)	0.22

MLE [42]	54	0.79	(0.66, 0.87)	0.21
Spearman	54	0.58	(0.34, 0.82)	0.48
Complete cases	15	0.76	(0.40, 0.92)	0.52
Kendall	54	0.49	(0.12, 0.74)	0.62

R code for applying all of the estimation methods considered by McCracken is freely available at <http://bcs.wiley.com/he-bcs/Books?action=resource&bcsId=9351&itemId=1118027558&resourceId=37125>.

$\hat{\rho}_s = 0.58$ with a 95% C.I. (ρ_s) of (0.34, 0.82). Thus, the association between *o*-cresol and hippuric acid appears to be quite a bit weaker than that claimed by Amorim and Alvarez-Leite ($r = 0.777$). The use of Spearman's coefficient with these data is consistent with Amorim and Alvarez-Leite's use of the nonparametric Kruskal-Wallis test to compare the *o*-cresol concentrations among the three groups of individuals exposed to toluene that they studied: workers in shoe factories, painting sectors of metal industries, and printing shops.

9. Discussion

Even if there are “no” data or data that appear to be of extremely limited usefulness, valid statistical methods are available, even when the sample size is quite small. Many of these methods are computationally simple; however, some of them require specialized software. We do not claim to have examined a comprehensive set of situations in which limited data may pose difficulties. Rather, we have focused on scenarios that we have encountered in scientific investigations in which we were directly involved. It is quite likely that other applied statisticians will have their own set of experiences with limited data that they can add to those we have described here.

References

- [1] Helsel, D. R. 2005. *Nondetects and Data Analysis*. Hoboken, NJ: John Wiley & Sons.
- [2] Helsel, D. R. 2010. “Much Ado about Next to Nothing: Incorporating Nondetects in Science.” *Annals of Occupational Hygiene* 54: 257-62.
- [3] Roberts, S., Vender, J. R., Causey, M. S., Roberts, J. R., Loushine, R. J., Morris, W. J., and Looney, S. W. 2009. “Impact of Dental Devices on Neurostimulators.” *Journal of Endodontics* 35: 422-5.
- [4] McCracken, C., and Looney, S. 2017. “On Finding the Upper Confidence Limit for a Binomial Proportion when Zero Successes Are Observed.” *Journal of Biometrics and Biostatistics* 8 (2): 338-43.
- [5] Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. 2nd ed., Hoboken, NJ: John Wiley & Sons.
- [6] Wilson, E. B. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22: 209-12.
- [7] Borkowf, C. B. 2006. “Constructing Binomial Confidence Intervals with Near Nominal Coverage by Adding a Single Imaginary Failure or Success.” *Statistics in Medicine* 25: 3679-95.
- [8] Tuyl, F., Gerlach, R., and Mengerson, K. 2008. “A Comparison of Bayes-Laplace, Jeffreys, and Other Priors: The Case of Zero Events.” *The American Statistician* 62: 40-4.
- [9] Clopper, C. J., and Pearson, E. S. 1934. “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial.” *Biometrika* 26: 404-13.
- [10] Leemis, L. M., and Trivedi, K. S. 1996. “A Comparison of Approximate Interval Estimators for the Bernoulli Parameter.” *The American Statistician* 50: 63-8.
- [11] Casella, G. 2001. “Comment on ‘Interval Estimation for a Binomial Proportion.’” *Statistical Science* 16: 120-2.
- [12] Agresti, A., and Coull, B. 1998. “Approximate Is Better than ‘Exact’ for Interval Estimation of Binomial Proportions.” *The American Statistician* 52: 119-26.
- [13] Vos, P., and Hudson, S. 2005. “Evaluation Criteria for Discrete Confidence Intervals: Beyond Length and Coverage.” *The American Statistician* 59: 137-42.
- [14] Meeks, S., Looney, S.W., Van Haitisma, K., and Teri, L. (2008). “BE-ACTIV: A Staff-Assisted, Behavioral Intervention for Depression in Nursing Homes.” *The Gerontologist* 48 (1): 105-14.
- [15] Agresti, A., and Caffo, B. 2000. “Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Result from Adding Two Successes and Two Failures.” *American Statistician* 54: 280-8.
- [16] Agresti, A., and Min, Y. 2001. “On Small-Sample Confidence Intervals for Parameters in Discrete

- Distributions.” *Biometrics* 57: 963-71.
- [17] Carter, R. E., Yan, L., Lipsitz, S. R., Newcombe, R. G., and Hermayer, K. L. 2010. “Relative Risk Estimated From the Ratio of Two Median Unbiased Estimates.” *Applied Statistics* 59: 657-71.
- [18] Blondell, R. D., Dodds, H. N., Looney, S. W., Lewis, C. M., Hagan, J. L., Lukan, J. K., and Servoss, T. J. 2005. “Toxicology Screening Results: Injury Associations Among Hospitalized Trauma Patients.” *Journal of Trauma* 58: 561-70.
- [19] Gart, J. J. 1970. “Point and Interval Estimation of the Common Odds Ratio in the Combination of 2×2 Tables with Fixed Marginals.” *Biometrika* 57 (3): 471-5.
- [20] London, S., Thomas, D., Bowman, J., Sobel, E., Cheng, T., and Peters, J. 1991. “Exposure to Residential Electric and Magnetic Fields and Risk of Childhood Leukemia.” *American Journal of Epidemiology* 134: 923-37.
- [21] Pike, M., Peters, R., Cozen, W., Probst-Hensch, N., Felix, J., Wan, P., and Mack, T. 1997. “Estrogen-Progestin Replacement Therapy with Endometrial Cancer.” *Journal of the National Cancer Institute* 89: 1110-6.
- [22] Haber, M., and Chen, C. 1991. “Estimation of Odds Ratios from Matched Case-Control Studies with Incomplete Data.” *Biometrical Journal* 33: 673-82.
- [23] Huberman, M., and Langholz, B. 1999. “Application of the Missing-Indicator Method in Matched Case-Control Studies with Incomplete Data.” *American Journal of Epidemiology* 150: 1340-5.
- [24] Miller, K. M., and Looney, S. W. 2012. “A Simple Method for Estimating the Odds Ratio in Matched Case-Control Studies with Incomplete Paired Data.” *Statistics in Medicine* 31 (27): 3299-312.
- [25] Hodges, J. L., and Lehmann, E. L. 1970. *Basic Concepts of Probability and Statistics*. 2nd ed., Oakland, CA: Holden-Day, 306-8.
- [26] Hermann, C., and Looney, S. 2001. “The Effectiveness of Symptom Management in Hospice Patients During the Last Seven Days of Life.” *Journal of Hospice and Palliative Care* 3: 88-96.
- [27] Dimery, I. W., Nishioka, K., Grossic, B., Ota, D., Schantz, S. P., Byers, R., Robbins, K. T., and Hong, W. K. 1987. “Polyamine Metabolism in Carcinoma of the Oral Cavity Compared with Adjacent and Normal Oral Mucosa.” *American Journal of Surgery* 154: 429-33.
- [28] Nurnberger, J., Jimerson, D. C., Allen, J. R., Simmons, S., and Gershon, E. 1982. “Red Cell Oabain-Sensitive Na⁺-K⁺-Adenosine Triphosphatase: A State Marker in Affective Disorder Inversely Related to Plasma Cortisol.” *Biological Psychiatry* 17: 981-92.
- [29] Steere, A. C., Green, J., Schoen, R. T., Taylor, E., Hutchinson, G. J., Rahn, D. W., and Malawista, S. E. 1985. “Successful Parenteral Penicillin Therapy of Established Lyme Arthritis.” *New England Journal of Medicine* 312: 869-74.
- [30] Snedecor, G. W., and Cochran, W. G. 1980. *Statistical Methods*. 7th ed., Ames, IA: Iowa State University Press, 101.
- [31] Looney, S. W., and Jones, P. W. 2003. “A Method for Comparing Two Normal Means Using Combined Samples of Correlated and Uncorrelated Data.” *Statistics in Medicine* 22: 1601-10.
- [32] Uddin, N., and Hasan, M. S. 2017. “Testing Equality of Two Normal Means Using Combined Samples of Paired and Unpaired Data.” *Communications in Statistics-Simulation and Computation* 46 (3): 2430-46.
- [33] Guo, B., and Yuan, Y. 2017. “A Comparative Review of Methods for Comparing Means Using Partially Paired Data.” *Statistical Methods in Medical Research* 26 (3): 1323-40.
- [34] Cohen, J. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20: 37-46.
- [35] Landis, J. R., and Koch, G. G. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33: 159-74.
- [36] Byrt, T., Bishop, J., and Carlin, J. B. 1993. “Bias, Prevalence, and Kappa.” *Journal of Clinical Epidemiology* 46: 423-9.
- [37] Looney, S. W., and Hagan, J. L. 2006. “On Methods for Handling Measurement Data below the Analytic Limit of Detection.” *Proceedings of the American Statistical Association*, 2477-81, presented at the 2006 Joint Statistical Meetings, Minneapolis, MN.
- [38] Amorim, L. C. A., and Alvarez-Leite, E. M. 1997. “Determination of *O*-Cresol by Gas Chromatography and Comparison with Hippuric Acid Levels in Urine Samples of Individuals Exposed to Toluene.” *Journal of Toxicology and Environmental Health* 50: 401-7.
- [39] Lagorio, S., Crebelli, R., Ricciarello, R., Conti, L., Iavarone, I., Zona, A., Ghittori, S., and Carere, A. 1998. “Methodological Issues in Biomonitoring of Low Level Exposure to Benzene.” *Occupational Medicine* 8: 497-504.
- [40] Atawodi, S. E., Lea, S., Nyberg, F., Mukeria, A., Constantinescu, V., Ahrens, W., Brueske-Hohlfeld, I., Fortes, C., Boffetta, P., and Friesen, M. D. 1998. “4-Hydroxyl-1-(3-pyridyl)-1-Butanone-Hemoglobin Adducts as Biomarkers of Exposure to Tobacco Smoke: Validation of a Method to Be Used in Multicenter Studies.” *Cancer Epidemiology Biomarkers and Prevention* 7: 817-21.
- [41] Cook, D. G., Whincup, P. H., Papacosta, O., Strachan, D. P., Jarvis, M. J., and Bryant, A. 1993. “Relation of Passive Smoking as Assessed by Salivary Cotinine

- Concentration and Questionnaire to Spirometric Indices in Children." *Thorax* 48: 14-20.
- [42] Lyles, R. H., Williams, J. K., and Chuachoowong, R. 2001. "Correlating Two Viral Load Assays with Known Detection Limits." *Biometrics* 57: 1238-44.
- [43] McCracken, C. E. 2013. "Correlation Coefficient Inference for Left-Censored Measurement Data with Known Detection Limits." unpublished Ph.D. dissertation, Georgia Regents University, Dept. of Biostatistics.
- [44] Lynn, H. 2001. "Maximum Likelihood Inference for Left-Censored HIV RNA Data." *Statistics in Medicine* 20: 33-45.
- [45] Li, L., Wang, W., and Chan, I. 2005. "Correlation Coefficient Inference on Censored Bioassay Data." *Journal of Biopharmaceutical Statistics* 15: 501-12.
- [46] Looney, S. W., and Hagan, J. L. 2015. *Analysis of Biomarker Data: A Practical Guide*. Hoboken, NJ: John Wiley & Sons.