J. Dairy Sci. 95:2576–2584
http://dx.doi.org/10.3168/jds.2011-4515

Technical note: The risk ratio, an alternative to the odds ratio for estimating the association between multiple risk factors and a dichotomous outcome

P. A. Ospina,* D. V. Nydam,†[1] and T. J. DiCiccio‡
*Department of Animal Science, College of Agriculture and Life Sciences,
†Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, and
‡Department of Social Statistics, School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14853

ABSTRACT

The objectives were (1) to explain why the risk ratio (RR) is an appropriate measure of association when the outcome of interest is dichotomous (e.g., displaced abomasum or no displaced abomasum) in both cohort studies and randomized trials; and (2) to outline an applied method for estimating the RR using currently available software. Interest in the association between multiple risk factors and a yes or no outcome is very common in the dairy industry; historically, logistic regression, which reports odds ratios (OR), was the method available in common statistical packages to evaluate this kind of association. However, the OR can overestimate the magnitude of the response in cohort studies and randomized trials when the outcome frequency is large. In addition, the interpretation of odds is not intuitive; fortunately, recent advances in statistical software have allowed the estimation of the RR. Because SAS software (SAS Institute Inc., Cary, NC) is commonly used to analyze data, this technical note outlines the basic programming code that may be used to estimate the RR from raw data. Example data from a prospective cohort study was used to compare the OR and RR of developing a displaced abomasum or ketosis or metritis based on multiple predictors, their interaction, and a random effect (e.g., herd).
Key words: risk ratio, odds ratio, Poisson, SAS software

Technical Note

Multivariable analysis; that is, evaluating the association between multiple predictor variables such as main effects, covariates, and potential confounders with an outcome of interest, is common in dairy science research. This kind of analysis with a yes or no outcome (e.g., mastitis or no mastitis) has become more common but estimating the magnitude of the effect has been limited to 2 × 2 tables, ordinary least square (**OLS**) linear regression, OLS with an arc sine transformation, or logistic regression.

Although odds ratios (**OR**) can be estimated in any study with a yes or no outcome, risk ratios (**RR**) can only be estimated in studies where probabilities are measured directly; that is, cohort studies and randomized trials. The calculation of an OR and an RR and the algebraic relationship between them is shown for a general 2 × 2 table in Table A1 of the Appendix. Although both measures of association can be estimated this way, when models contain several covariates and interaction terms, this approach is limiting and laborious.

Although OLS linear regression can be used for analysis of data with several covariates, interaction terms, and a dichotomous outcome, several restrictions and assumptions are violated when the outcome is dichotomous (Dohoo et al., 2003) and the probability of success is <0.1 or >0.9 (Cox, 1970). The most severe consequence results from the violation of the homoscedasticity assumption. This violation can lead to inefficient estimates of coefficients and biased standard error estimates, which result in biased test statistics (Allison, 2001).

An alternative approach to multivariable modeling of positive dichotomous outcomes has been logistic regression, which is based on the logit-transformation and maximum likelihood estimation. This transformation replaces the probability of a positive dichotomous outcome with the log odds, and re-establishes the linear link between this and the predictor variables.

A search of the *Journal of Dairy Science* online through April 6, 2011, revealed that the first article to discuss the use of logistic regression for the analysis of a dichotomous outcome dates to 1987 (Curtis et al., 1987). Since then, 370 articles published in the *Journal of Dairy Science* have used the exact phrase "logistic regression" in the title, abstract, or text. The OR was

Received May 6, 2011.
Accepted November 17, 2011.
[1]Corresponding author: dvn2@cornell.edu

frequently used because it is the estimable measure of association in some study designs (i.e., case-control studies) or because estimating a more appropriate measure of association such as the RR was not readily feasible with commercially available statistical packages. Although the OR and not the RR was estimated, instead of reporting the odds, the results were sometimes misinterpreted as RR and incorrect phrases such as "more likely" or "risk" were used to describe the association between the risk factor(s) and the outcome of interest.

Two problems are associated with estimating the OR in study designs where the RR can be estimated: (1) the OR can overestimate the true effect as the outcome becomes more common, and (2) interpretation of the OR is not intuitive (Holcomb et al., 2001) because, unlike the RR, the OR does not directly measure effects on probability. The objectives of this study are to explain why the RR is the preferred measure of association when the outcome of interest is dichotomous in both cohort studies and randomized trials and to outline an applied method for estimating the RR using SAS version 9.2 (SAS Institute Inc., Cary, NC).

Objective 1: Preference of RR to OR in Certain Study Designs

Although estimating the OR is not incorrect, in study designs where the RR can be estimated it is considered the more appropriate measure of association (Greenland, 1987; Spiegelman and Hertzmark, 2005), because it provides estimates of probabilities directly. The OR can overestimate the true effect when interpreted as a probability in studies where the frequency is large, and the interpretation of odds is not as intuitive as probabilities. To demonstrate that the OR can overestimate the RR as the incidence of the outcome increases, simple example calculations of OR versus RR are provided in Table A2 of the Appendix. The incidence ranges from 5 to 30%. The example with 30% incidence could reflect that seen in a study of mastitis and demonstrates the potential overestimation of the effect when an OR is estimated instead of an RR.

The RR can be calculated from cohort studies and cross-sectional studies as well as randomized trials. These are subsets of 2 general study designs, observational and experimental, respectively. When an experiment is not feasible, observational studies are used to evaluate whether an association exists between a risk factor and an outcome of interest. These studies can be divided into 3 groups: cross-sectional, cohort, and case-control. In a cross-sectional study, both exposure and the outcome of interest are determined simultaneously; therefore, information is lacking about whether the exposure preceded the outcome. In a cohort study, subjects are selected based on whether they have been exposed to the risk factor or not and then followed to determine whether they develop the outcome of interest; thus, outcome probabilities can be estimated directly. Conversely, in a case-control study, subjects are selected based on whether they have the outcome of interest (case) or not (control) and then information on whether they were exposed to the risk factor is collected retrospectively. This type of selection is used to investigate rare outcomes; however, it results in a predetermined ratio of cases to controls and only the ratio of the odds can be estimated.

In situations where an experiment is feasible, a randomized trial is considered the best method to determine causality between a risk factor (e.g., treatment) and the outcome of interest. Randomization reduces potential confounding and more control exists over exposure to the risk factor of interest. For example, to determine whether treatment with propylene glycol (exposure) to transition cows reduces the risk of the development of ketosis (outcome), cows can be randomly assigned to receive propylene glycol or placebo, and then the RR or OR of developing ketosis given administration of propylene glycol can be estimated.

In the preceding example, both the OR and the RR could be estimated. However, the OR reports the ratio of the odds that the cow will develop ketosis, given that she was treated with propylene glycol, to the odds that the cow will develop ketosis, given that she was not treated, whereas the RR reports the ratio of the probability that the cow will develop ketosis, following exposure, to the probability that the cow will develop ketosis when not exposed. Because this was a randomized trial and the RR was estimable, the OR serves only as a potential overapproximation of the RR.

Objective 2: Calculating the RR with Poisson Regression

In light of some of the concerns with the use of the OR in cohort studies (Greenland, 1987) and randomized trials, several methods have been proposed to estimate the RR. The use of a log link with a binary outcome appears to be a good choice because it can calculate the RR directly. However, this method can have issues with convergence because the estimated probabilities are not confined to the allowable parameter space between 0 and 1 (McNutt et al., 2003). Although several techniques to address the convergence issues have been proposed, they can be complex and difficult to implement (Zocchetti et al., 1995). Other authors

(Zhang and Yu, 1998; Kleinman and Norton, 2009) advocated the use of logistic regression to estimate the RR through additional calculations, using the fact that the odds is equal to [P/(1 − P)], where P = probability, as a way to avoid the parameter space limitations and thus, avoid convergence issues; however, this can lead to biased estimators (McNutt et al., 2003; Zou, 2004). Poisson regression has its own limitations, namely, that it might provide conservative results (Zou, 2004); that is, the confidence interval will be wider. It can have convergence issues due to parameter space concerns as in the log binomial model, but this is rare in practice and the model will not yield results if it does happen. Poisson regression has several advantages such as, when used with binary data, the estimates are unbiased, which means that Poisson regression will produce consistent estimations of the relative risk (Lumley et al., 2006). Additionally, when used in cohort studies, it can account for different times at risk (Zou, 2004), a function that cannot be supported by the log-binomial model.

The use of Poisson regression to calculate RR has been advocated by several authors (Frome and Checkoway, 1985; McNutt et al., 2003; Zou, 2004). Although Poisson regression is commonly associated with the evaluation of count data, it can also be used for binary data. The only model restrictions are as follows: the dependent variable must be a nonnegative integer and it must have a Poisson distribution conditional on the values of the explanatory variables (Allison, 2001). Data from a prospective cohort study of 1,318 Holsteins in 100 herds (Ospina et al., 2010) were used to demonstrate the method for calculating the RR with PROC GENMOD using SAS. In this study, approximately 15 cows in the postpartum period (3 to 14 DIM) were sampled per herd. The sampling consisted of drawing blood and measuring a single serum BHBA concentration from healthy cows and heifers. These animals were followed to 30 DIM, and within this timeframe, 3 disease outcomes were documented: (1) displaced abomasum, (2) ketosis, or (3) metritis. The dichotomous outcome of interest was the presence or absence of any of the 3 diseases. The predictor variables were (1) exposure to an elevated concentration of BHBA between 3 to 14 DIM (main effect), (2) parity (potential confounder), (3) the interaction between BHBA and parity, and (4) herd as a random effect. The concentration at which BHBA was considered elevated was 10 mg/dL, which was previously determined by receiver operator characteristic curve analysis (Ospina et al., 2010). Parity was dichotomized as 1 or ≥2. The SAS code without and with the interaction term will be demonstrated and discussed in the next two sections.

The RR (without interaction term) can be estimated with PROC GENMOD, as follows:

```
proc genmod descending data = work.
BHBAstudy;
class herd BHBA Parity;
model disease = BHBA Parity / link = log dist
= Poisson pscale type3;
repeated subject = herd/ type = exch;
estimate 'BHBA' BHBA 1 -1/ exp;
estimate 'Parity' Parity 1 -1/exp;
run;
```

The first line of the SAS code tells the system what procedure to run, how to read the data, and where to find the data. The procedure is GENMOD and, in this case, the data can be found in a work file titled "BHBAstudy." The word "descending" relates to how the outcome data are read. By default, when evaluating a dichotomous outcome, the program will model the probability of the smallest outcome variable. In most cases, the outcome variable is coded as 0 or 1, where 0 means that the outcome of interest (event) did not occur, and 1 means that the outcome of interest (event) did occur. Research interest usually lies in evaluating the risk of developing the outcome of interest, so adding the "descending" option to the first line tells the program to model the probability of the largest outcome variable.

The second line of the code is the class statement. This tells the program that the variables "herd," "BHBA," and "parity" are categorical variables. Any class variables, including those in the repeated statement should be included in this line.

The third line is the model statement. The model statement defines the equation that will be examined. In this case, the association between the outcome of interest (occurrence of any of the 3 diseases) and the predictor variables (main effects, exposure to an elevated concentration of BHBA; potential confounders, parity) will be evaluated. To run Poisson regression using PROC GENMOD, the log link and Poisson distribution must be specified after a forward slash (/) in the model statement.

A major assumption of the Poisson distribution is that the variance is equal to the mean (Agresti, 2007); however, unless all factors are controlled for in a model, this is rarely the case and overdispersion usually occurs. Overdispersion does not bias the coefficients, but it can underestimate the standard errors (Allison, 2001). Correcting the standard errors for overdispersion is done by adding the "pscale" option to the model statement. Under the "pscale" option, each standard error is multiplied by the square root of the Pearson chi-squared

Table 1. Output from SAS software (SAS Institute Inc., Cary, NC) for estimating the risk ratio (RR) of developing a displaced abomasum, ketosis, or metritis with Poisson regression using PROC GENMOD

Analysis of generalized estimating equations (GEE) parameter estimates

Empirical SE estimates

Parameter		Estimate	SE	95% CI		Z	Pr > |Z|
Intercept		−3.0450	0.2097	−3.4560	−2.6341	−14.52	<0.0001
BHBA[1]	1	1.4946	0.1828	1.1363	1.8529	8.18	<0.0001
BHBA	10	0.0000	0.0000	0.0000	0.0000	—	—
Parity[2]	1	0.0350	0.1738	−0.3057	0.3757	0.2	0.8406
Parity	2	0.0000	0.0000	0.0000	0.0000	—	—

Score statistics for type 3 GEE analysis

Source	df	χ^2	Pr > χ^2
BHBA	1	29.32	<0.0001
Parity	1	0.04	0.8437

Contrast estimate results

Label	Estimate	SE	α	95% CI		χ^2	Pr > χ^2
BHBA	1.4946	0.1828	0.05	1.1363	1.8529	66.84	<0.0001
Exp (BHBA)	4.4576[3]	0.8149	0.05	3.1152	6.3785		
Parity	0.0350	0.1738	0.05	−0.3057	0.3757	0.04	0.8406
Exp (Parity)	1.0356[4]	0.1800	0.05	0.7366	1.4560		

[1]BHBA dichotomized: 10 if < 10mg/dL (reference level) and 1 if ≥10 mg/dL.

[2]Parity dichotomized: 1 if parity = 1 and 2 if ≥2 (reference level).

[3]Risk ratio of developing any of the 3 diseases if BHBA concentration was ≥10 mg/dL.

[4]Risk ratio of developing any of the 3 diseases if parity was = 1.

statistic for testing goodness of fit, and divided by its degrees of freedom. The "pscale" option is preferred to the "dscale" option due to the theory of quasi-likelihood estimation (McCullagh and Nelder, 1989). The "type3" option offers the opportunity to examine the results without regard to the order in which the terms are specified in the model statement and analyses are based on single degree of freedom. Although these score statistics evaluate the same hypothesis as the Z-statistics in the generalized estimating equation (GEE) parameter estimates, the score statistics usually report more conservative *P*-values (Stokes et al., 2000). Reporting these *P*-values should be considered in studies with small sample sizes.

The fourth line of code is the repeated statement. In PROC GENMOD, a class variable (e.g., herd) can be used to cluster individual samples (e.g., cows). It is reasonable to suspect that more similarities will exist among cows within the same herd versus those between herds (McDermott et al., 1994a,b). In addition to the measured characteristics included in the model, many unmeasured characteristics in the herd will likely have an effect on the probability of whether a cow gets any of the 3 diseases, for example. These intra-herd similarities can be taken into account by incorporating the repeated-measures statement and specifying the structure of the correlation matrix as "exch." Choosing the "exch" option specifies a single correlation that applies to any pair of animals within each cluster (Allison, 2001), meaning that within each herd, the probability of developing the outcome of interest should be similar given levels of the risk factors, but this probability can be different between herds. In addition, this modification can be used to estimate correctly the standard error for the RR (Zou, 2004).

The fifth and sixth lines of code are the estimate statements. These statements serve 2 functions: allow contrasts and exponentiate the estimate from the base *e* ("exp" option after the forward slash). This exponentiation facilitates evaluation of the RR because it no longer needs to be done by hand. It is important to examine the estimate statement closely as this will determine the interpretation of the results. The "1 −1" in this statement compares the BHBA concentration ≥10 mg/dL to BHBA <10 mg/dL (the reference level) and estimates the RR with a 95% CI based on the Wald statistic. It is important to note that SAS will choose the predictor variable with the largest numerical value as the reference level. Estimate statements can be written for any variable in the model. The SAS output generated from running this code can be found in Table 1.

As shown in Table 1, the RR for developing any of the 3 diseases given that BHBA was ≥10 mg/dL was 4.5, meaning that cows with BHBA concentration ≥10 mg/dL were 4.5 times more likely to develop a displaced abomasum, clinical ketosis, or metritis than cows with BHBA <10 mg/dL. However, the OR estimated using

PROC LOGISTIC was a slight overestimation at 5.1. The interpretation of the OR is that a cow with BHBA ≥10 mg/dL has 5.1 greater odds of developing any of the 3 diseases than a cow with BHBA <10 mg/dL. Note that this is not the likelihood or probability of having the event, but specifically the odds.

The RR (with interaction term) can be calculated with PROC GENMOD, as follows:

```
proc genmod descending data = work.
BHBAstudy;
class herd BHBA Parity;
model Disease = BHBA Parity BHBA*Parity /
link = log dist = Poisson pscale type3;
repeated subject = herd/ type = exch;
lsmeans BHBA10 * Parity / pdiff cl exp;
run;
```

Evaluating a 1-way interaction is useful for testing the hypothesis that the relationship between one predictor variable and the outcome depends on the level of the second predictor variable. In this example, the interaction between BHBA and parity was evaluated to see if the effect of elevated BHBA on the development of any of the 3 diseases was different if parity = 1 or if parity ≥2, or if the effect of parity on the development of disease was different if BHBA was ≥10 mg/dL or <10 mg/dL.

A few key differences can be seen in this code compared with the code without the interaction term. The first major difference can be found in the third line, the model statement. The model statement now contains an interaction term, "BHBA * Parity." Interaction terms are incorporated into the model by taking the individual variables and adding an asterisk between them; that is, var 1 * var 2 (e.g., BHBA * Parity). Because the BHBA example consisted of categorical variables, a discussion of continuous variables is not included; however, the interpretation of interactions with categorical variables is similar to that of continuous variables.

The "lsmeans" statement with "pdiff," "cl," and "exp" options is the second major difference in this code. This "lsmeans" statement allows the examination of all the permutations of categorical variables in the interaction terms, and one "lsmeans" statement should be written for each interaction term in the model. The output generated from this statement will give the RR for all of the permutations and the "cl" option provides 95% confidence limits. Although 95% is standard, this can be changed by suggesting a different α; to do so, add "alpha = #" after the "cl" term. This will also work under the "estimate" statement in the previous example. Generally, when an interaction term is included in a model statement there is no longer one single estimate that determines the RR of the outcome based on the terms included in the interaction term; it is now dependent on the level of the second variable.

The SAS output generated from running the model statement can be found in Tables 2, 3, and 4. Table 2 contains the information from running the model statement, the output from the LSMEANS statement is in Table 3, and the outputs from the PDIFF, CL, and exp options are found in Table 4. The inclusion of the interaction term in the model statement means that the single estimate listed in Table 2 of 1.28 for BHBA can no longer solely be used to estimate the RR of developing any of the 3 diseases, because the effect of BHBA on disease risk depends on the level of parity. Also, in Table 2, the estimate of 0.60 for the interaction term (BHBA * Parity) represents the change in the estimate of the main effect (BHBA) when the other variable (parity) changes by 1 level. The *P*-value associated with the interaction term at this level of analysis is evaluating whether the interaction has a statistically significant effect on the outcome.

The information in Table 3 offers the opportunity to examine the risks of developing the outcome (any of the 3 diseases). See Table A3 in the Appendix for the mathematical formulation used to evaluate the

Table 2. Output from SAS software (SAS Institute Inc., Cary, NC) for estimating the risk ratio (RR) with Poisson regression using PROC GENMOD to evaluate the model[1] with an interaction term

Analysis of generalized estimating equation parameter estimates (empirical SE estimates)					
Parameter	Estimate	SE	95% CI		Pr > |Z|
Intercept	−2.9157	0.2092	−3.3257	−2.5057	<0.0001
BHBA[2]	1.2844	0.2072	0.8783	1.6905	<0.0001
Parity[3]	−0.3265	0.2953	−0.9053	0.2523	0.2689
BHBA * Parity	0.6008	0.3447	−0.0749	1.2765	0.0814

[1]The model: disease outcomes (displaced abomasum, or ketosis, or metritis) = BHBA[2] + Parity[3] + BHBA * Parity + cows clustered within herds.

[2]BHBA dichotomized: 10 if <10 mg/dL (reference level) and 1 if ≥10 mg/dL.

[3]Parity dichotomized: 1 if parity = 1 and 2 if ≥2 (reference level).

Table 3. Abbreviated SAS (SAS Institute Inc., Cary, NC) output generated from LSMEANS statement[1]

BHBA * Parity least squares means					
BHBA[1]	Parity[2]	Estimate	SE	Z-value	Pr > |Z|
1	1	−1.3569	0.1658	−8.18	<0.0001
1	2	−1.6313	0.1501	−10.86	<0.0001
0	1	−3.2422	0.2672	−12.14	<0.0001
0	2	−2.9157	0.2092	−13.94	<0.0001

[1]The following columns were omitted: α, lower and upper confidence limits for estimate, exponentiated estimate, and 95% CI of exponentiated "estimate."

interaction term. To calculate the RR, the information in Table 4 or the calculations in the Appendix are necessary. The information in Table A3 shows how the numbers in Table 3 were estimated; that is, where they came from, and shows simple calculations of the RR when an interaction term is included. Table 4 allows independent evaluation of each of the permutations of the interactions and tests whether the difference is significantly different from zero. The column labeled "exponentiated" has the RR for each of the permutations of the interaction term. The confidence limits for the RR are estimated in Table 4. The Appendix has sample calculations of RR of interest. Because a log link was used, the natural log is taken on both sides of the equation to evaluate the estimates. In this example, the RR for the development of diseases given that BHBA was ≥10 mg/dL was 6.6 when parity was = 1, and 3.6 when parity was ≥2.

Because incorporation of an interaction term can make interpretation of the model more complex, the decision to include an interaction term in a model should not be taken lightly. As previously stated, Table 4 allows the evaluation of all permutations of the interaction model and in this case, no significant difference ($P = 0.17$, 0.27) was found when permutations were compared across parity levels and BHBA levels were held constant. This information along with the large likelihood that the difference based on parity was due to chance ($P = 0.2689$) may result in deciding to remove parity and the interaction term from the model.

Poisson regression was used to estimate the RR. Poisson regression analysis is appropriate when the response variable is a count or a rate and the results can be interpreted as a rate ratio (Frome and Checkoway, 1985). Counts represent the number of events that occurred over an observed period; however, if that period is similar between all subjects, then the rate ratio approximates the RR. This relationship can be seen with the following equations of a hypothetical example when each cow contributes 1 cow-week of time at risk:

$$\text{Rate ratio} = \frac{\text{Incidence rate in exposed}}{\text{Incidence rate in unexposed}} = \frac{3 \text{ cases} / 10 \text{ cow-weeks}}{1 \text{ case} / 10 \text{ cow-weeks}};$$

$$\text{Risk ratio} = \frac{\text{Incidence risk in exposed}}{\text{Incidence risk in unexposed}} = \frac{3 \text{ cases} / 10 \text{ exposed cows}}{1 \text{ case} / 10 \text{ exposed cows}}.$$

It should be noted that if some counts correspond to different periods at risk than others, then the rate at

Table 4. Abbreviated SAS (SAS Institute Inc., Cary, NC) obtained from the "PDIFF CL Exp" options in the LSMEANS statement[1]

Differences of least square means									
BHBA	Parity	BHBA	Parity	Estimate	SE	*P*-value[2]	Exponen.[3] lower	Exponen.[4] upper	Exponentiated[5]
1	1	1	2	0.2743	0.1979	0.1658	0.8926	1.9392	1.3156
1	1	0	1	1.8852	0.3009	<0.0001	3.6526	11.8817	6.5878
1	1	0	2	1.5587	0.2418	<0.0001	2.9588	7.6348	4.7529
1	2	0	1	1.6109	0.2879	<0.0001	2.8514	8.7932	5.0073
1	2	0	2	1.2844	0.2072	<0.0001	2.4069	5.4223	3.6126
0	1	0	2	−0.3265	0.2953	0.2689	0.4044	1.2870	0.7215

[1]The following columns were omitted: Z-value, α, and lower and upper confidence limits for estimate.

[2]*P*-value for the hypothesis that the difference between the 2 levels within the interaction was different from zero.

[3]Lower 95% confidence limit of exponentiated "estimate"; that is, the lower 95% CI for risk ratio for specific interaction.

[4]Upper 95% confidence limit of exponentiated "estimate"; that is, the upper 95% CI for risk ratio for specific interaction.

[5]The risk ratio for each permutation of the interaction term.

which the events occur must be included in the model (Stokes et al., 2000; Zou, 2004). This is done by including an "offset" term after the forward slash in the model statement (e.g., offset = time at risk). The natural log of the time at risk should be calculated to include this information in the model. In the current BHBA study, all animals were at risk for similar periods, so the offset term was not included (i.e., was equal to 1).

Interpretation of an RR is intuitive. Many studies that estimate OR interpret it incorrectly as an RR (Holcomb et al., 2001). It is important to note that when the incidence of the outcome is low (e.g., $\leq 5\%$), the RR and OR are very similar in magnitude. Yet, the difference between the estimate of the OR versus the RR becomes larger and therefore more of a concern as the incidence becomes greater (Appendix Tables A1 and A2). The RR reports the likelihood that the outcome of interest will happen, given a certain level of the risk factor, and is based on the idea that the sample for this calculation is representative of the true population. This SAS code will estimate the RR and therefore, eliminates the need to estimate the OR in cohort studies and randomized trials.

REFERENCES

Agresti, A. 2007. An Introduction to Categorical Data Analysis. Wiley-Interscience, Hoboken, NJ.

Allison, P. D. 2001. Logistic Regression using the SAS System: Theory and Application. SAS Institute Inc., Cary, NC.

Cox, D. R. 1970. Analysis of Binary Data: Monographs on Statistics and Applied Probability. Cambridge University Press, Cambridge, UK.

Curtis, C. R., H. N. Erb, and J. M. Kranz. 1987. An introduction to the use of epidemiologic research methods in dairy science. J. Dairy Sci. 70:373–380.

Dohoo, I. R., W. Martin, and H. Stryhn. 2003. Logistic regression. Veterinary Epidemiologic Research. 1st ed. AVC Inc., Charlottetown, PEI, Canada.

Frome, E. L., and H. Checkoway. 1985. Epidemiologic programs for computers and calculators. use of poisson regression models in estimating incidence rates and ratios. Am. J. Epidemiol. 121:309–323.

Greenland, S. 1987. Interpretation and choice of effect measures in epidemiologic analyses. Am. J. Epidemiol. 125:761–768.

Holcomb, W. L. Jr., T. Chaiworapongsa, D. A. Luke, and K. D. Burgdorf. 2001. An odd measure of risk: Use and misuse of the odds ratio. Obstet. Gynecol. 98:685–688.

Kleinman, L. C., and E. C. Norton. 2009. What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. Health Serv. Res. 44:288–302.

Lumley, T., R. Kronmal, and M. Shuangge. 2006. UW Biostatistics Working Paper Series 2006 (paper no. 293): Relative risk regression in medical research: Models, contrasts, estimators and algorithms. Accessed Feb. 13, 2012. http://www.bepress.com/uwbiostat/paper293.

McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. 2nd ed. Chapman and Hall, London, UK.

McDermott, J. J., Y. H. Schukken, and M. Shoukri. 1994a. A review of methods used to adjust for cluster effects in explanatory epidemiological studies of animal populations. Prev. Vet. Med. 18:155–173.

McDermott, J. J., Y. H. Schukken, and M. Shoukri. 1994b. Study design and analytic methods for data collected from clusters of animals. Prev. Vet. Med. 18:175–191.

McNutt, L. A., C. Wu, X. Xue, and J. P. Hafner. 2003. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am. J. Epidemiol. 157:940–943.

Ospina, P. A., D. V. Nydam, T. Stokol, and T. R. Overton. 2010. Evaluation of nonesterified fatty acids and β-hydroxybutyrate in transition dairy cattle in the northeast USA: Critical thresholds for prediction of clinical diseases. J. Dairy Sci. 93:546–554.

Spiegelman, D., and E. Hertzmark. 2005. Easy SAS calculations for risk or prevalence ratios and differences. Am. J. Epidemiol. 162:199–200.

Stokes, M. E., D. S. Davis, and G. G. Koch. 2000. Categorical Data Analysis Using the SAS System. 2nd ed. SAS Institute Inc., Cary, NC.

Zhang, J., and K. F. Yu. 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 280:1690–1691.

Zocchetti, C., D. Consonni, and P. A. Bertazzi. 1995. Estimation of prevalence rate ratios from cross-sectional data. Int. J. Epidemiol. 24:1064–1067.

Zou, G. 2004. A modified Poisson regression approach to prospective studies with binary data. Am. J. Epidemiol. 159:702–706.

APPENDIX

In a case control study (based on the 2 × 2 table in Table A1), the odds ratio (**OR**) is calculated as follows:

$$\frac{\text{odds that a case was exposed to the risk factor}}{\text{odds that a control was exposed to the risk factor}} = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

Table A1. An example of a 2 × 2 table

	Outcome (positive)[1]	Outcome (negative)[2]	Total
Exposed to risk factor	a	b	$a + b$
Not exposed to risk factor	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

[1]Positive outcome = individual developed the outcome of interest

[2]Negative outcome = individual did not develop the outcome of interest.

In a cohort study or randomized trial, the risk ratio (**RR**) is calculated as follows:

$$\frac{\text{incidence of outcome in those exposed to the risk factor}}{\text{incidence of outcome in those not exposed to the risk factor}} = \frac{a/\left(a+b\right)}{c/\left(c+d\right)}.$$

Thus (when a and c are relatively small), the RR approximates the OR as follows:

$$\frac{a/\left(a+b\right)}{c/\left(c+d\right)} = \frac{a\left(c+d\right)}{c\left(a+b\right)} \sim \frac{ad}{bc}.$$

Table A2. Five sample calculations of OR and RR[1]

Example 1. Incidence of the outcome was 2.5% and χ^2 *P*-value testing a difference between the exposed and not exposed group was 0.002; the OR and RR were similar (4.1 vs. 4, respectively).

Incidence 2.5%	Outcome (+)	Outcome (−)	Total	Results	
Exposed (+)	20	480	500	OR	4.1
Exposed (−)	5	495	500	RR	4.0
Total	25	975	1,000	χ^2	$P = 0.002$

Example 2. Incidence of the outcome was 5% and χ^2 *P*-value testing a difference between the exposed and not exposed group was <0.001; the OR was larger than RR (4.3 vs. 4, respectively).

Incidence 5%	Outcome (+)	Outcome (−)	Total	Results	
Exposed (+)	40	460	500	OR	4.3
Exposed (−)	10	490	500	RR	4.0
Total	50	950	1,000	χ^2	$P < 0.001$

Example 3. Incidence of outcome was 10% and χ^2 *P*-value testing a difference between the exposed and not exposed group was <0.001; the OR was larger than RR (4.6 vs. 4, respectively).

Incidence 10%	Outcome (+)	Outcome (−)	Total	Results	
Exposed (+)	80	420	500	OR	4.6
Exposed (−)	20	480	500	RR	4.0
Total	100	900	1,000	χ^2	$P < 0.001$

Example 4. Incidence of outcome was 15% and χ^2 *P*-value testing a difference between the exposed and not exposed group was <0.001; the OR was larger than RR (5 vs. 4, respectively).

Incidence 15%	Outcome (+)	Outcome (−)	Total	Results	
Exposed (+)	120	380	500	OR	5.0
Exposed (−)	30	470	500	RR	4.0
Total	150	850	1,000	χ^2	$P < 0.001$

Example 5. Incidence of outcome was 30% and χ^2 *P*-value testing a difference between the exposed and not exposed group was <0.001; the OR was larger than RR (6.8 vs. 4, respectively).

Incidence 30%	Outcome (+)	Outcome (−)	Total	Results	
Exposed (+)	240	260	500	OR	6.8
Exposed (−)	60	440	500	RR	4
Total	300	700	1,000	χ^2	$P < 0.001$

[1]Examples have outcome incidences that range from 2.5 to 30%. This is used to illustrate the point that the discrepancy between OR and RR gets larger as the incidence increases.

Table A3. Calculations of the risk of developing any of the 3 diseases and example calculations[1] of the RR with the incorporation of an interaction term

Model[1]: $Y = B_0 + B_1 \times$ BHBA (0,1) $+ B_2 \times$ Parity (0,1) $+ B_3 \times$ BHBA (0,1) $\times$ Parity (0,1)						
Y =	B_0	$B_1 \times$ BHBA	$B_2 \times$ Parity	$B_3 \times$ BHBA $\times$ Parity	Risk of Y[2]	Given:
DA, CK, or MET	− 2.9	+ 1.28 × 1	− 0.3 × 1	+ 0.6 × 1 × 1	$e^{-1.36}$	BHBA ≥10 mg/dL Parity = 1
	− 2.9	+ 1.28 × 1	− 0.3 × 0	+ 0.6 × 1 × 0	$e^{-1.63}$	BHBA ≥10 mg/dL Parity ≥2
	− 2.9	+ 1.28 × 0	− 0.3 × 1	+ 0.6 × 0 × 1	$e^{-3.24}$	BHBA <10 mg/dL Parity = 1
	− 2.9	+ 1.28 × 0	− 0.3 × 0	+ 0.6 × 0 × 0	$e^{-2.92}$	BHBA <10 mg/dL Parity ≥2

[1]Y = any combination of the 3 diseases: displaced abomasum (DA), clinical ketosis (CK), or metritis (MET); BHBA dichotomized: 0 if <10 mg/dL (reference level) and 1 if ≥10 mg/dL; Parity dichotomized: 1 if parity = 1 and 0 if ≥2 (reference level); $B_3 = 0.6$; this is the change in the estimate of the main effect (BHBA) when the other variable (parity) changes by 1 level.

[2]The exponent calculated here is the estimate in the "estimate" column in Table 3.

Based on the information in Table A3, the following are examples of how the RR (reported as "Exponentiated[5]" in Table 4) were derived:

$$\frac{e^{B_0+B_1(1)+B_2(1)+B_3(1*1)}}{e^{B_0+B_1(0)+B_2(1)+B_3(0*1)}} = \frac{e^{B_1(1)+B_3(1*1)}}{e^{B_1(0)+B_3(0*1)}} = \frac{e^{1.28+0.6}}{e^{0+0}=1} = e^{1.88} = 6.6.$$

The 6.6 value can be interpreted as the RR of developing any of the 3 diseases among animals in lactation 1 when BHBA was ≥10 mg/dL, relative to animals in the same parity with lower BHBA.

$$\frac{e^{B_0+B_1(1)+B_2(0)+B_3(1*0)}}{e^{B_0+B_1(0)+B_2(0)+B_3(0*0)}} = \frac{e^{B_1(1)+B_3(1*1)}}{e^{B_1(0)+B_3(0*1)}} = \frac{e^{1.28}}{e^{0}=1} = e^{1.28} = 3.6.$$

The 3.6 value can be interpreted as the RR of developing any of the 3 diseases among animals in lactation 2 when BHBA was ≥10 mg/dL, relative to animals in the same parity with lower BHBA.