

## Comparison Between a Weibull Proportional Hazards Model and a Linear Model for Predicting the Genetic Merit of US Jersey Sires for Daughter Longevity

D. Z. Caraviello, K. A. Weigel, and D. Gianola

Department of Dairy Science,  
University of Wisconsin, Madison 53706

### ABSTRACT

Predicted transmitting abilities (PTA) of US Jersey sires for daughter longevity were calculated using a Weibull proportional hazards sire model and compared with predictions from a conventional linear animal model. Culling data from 268,008 Jersey cows with first calving from 1981 to 2000 were used. The proportional hazards model included time-dependent effects of herd-year-season contemporary group and parity by stage of lactation interaction, as well as time-independent effects of sire and age at first calving. Sire variances and parameters of the Weibull distribution were estimated, providing heritability estimates of 4.7% on the log scale and 18.0% on the original scale. The PTA of each sire was expressed as the expected risk of culling relative to daughters of an average sire. Risk ratios (RR) ranged from 0.7 to 1.3, indicating that the risk of culling for daughters of the best sires was 30% lower than for daughters of average sires and nearly 50% lower than for daughters of the poorest sires. Sire PTA from the proportional hazards model were compared with PTA from a linear model similar to that used for routine national genetic evaluation of length of productive life (PL) using cross-validation in independent samples of herds. Models were compared using logistic regression of daughters' stayability to second, third, fourth, or fifth lactation on their sires' PTA values, with alternative approaches for weighting the contribution of each sire. Models were also compared using logistic regression of daughters' stayability to 36, 48, 60, 72, and 84 mo of life. The proportional hazards model generally yielded more accurate predictions according to these criteria, but differences in predictive ability between methods were smaller when using a Kullback-Leibler distance than with other approaches. Results of this study suggest that survival analysis methodology may provide

more accurate predictions of genetic merit for longevity than conventional linear models.

**(Key words:** longevity, survival analysis, proportional hazards, dairy cattle)

**Abbreviation key:** **KL** = Kullback-Leibler, **PL** = productive life, **RR** = relative risk.

### INTRODUCTION

New opportunities for genetic improvement of longevity or length of productive life (**PL**) in dairy cattle have recently become available because of modern, powerful, computer systems and widespread availability of software packages that allow application of advanced statistical methodology to large data sets. Survival or failure time analysis has replaced linear model approaches for routine genetic evaluation of dairy sires in several countries. Weibull proportional hazards models (Ducrocq and Solkner, 1998a) are generally used, and these provide heritability estimates for longevity from 0.15 to 0.20 as compared with 0.05 to 0.10 from conventional linear models. To the extent that such differences in heritability translate into more rapid genetic progress, implementation of survival analysis methodology could have an important impact on dairy cattle breeding programs.

In the US, a linear model is used for genetic evaluation of PL, as described by VanRaden and Klaaskate (1993). For completed records, PL is calculated as the total number of months in milk between first calving and 84 mo of age, with a maximum of 10 mo per lactation. For censored records from cows that were sold for dairy purposes or cows that are still alive at the time of analysis, linear regression is used to "project" total months in milk from data regarding current months in milk, current months dry, age at first calving, and lactation status (milking or dry). However, the correlation between completed PL records and their early projections is low (VanRaden and Klaaskate, 1993), so the reliability of information for sires that have the majority of their daughters in first and second lactation is limited.

---

Received October 9, 2003.

Accepted January 6, 2004.

Corresponding author: K. A. Weigel; e-mail: weigel@calshp.cals.wisc.edu.

Survival analysis is based on the concept of the hazard rate, which is the instantaneous probability of culling for a particular animal at a given point in time (Smith and Quaas, 1984; Ducrocq, 1987). The application of survival analysis methodology does not require extra data collection; it is simply an improvement in the statistical treatment of culling data.

Censored observations can be accommodated properly in survival models because one can differentiate between a cow that died at exactly time  $t$  (i.e., a completed record) and a cow that was last seen alive at time  $t$  but may have survived several additional months or years (i.e., a censored record). Survival models can also handle time-dependent covariates, and this feature is particularly important when factors such as herd size, facilities, management practices, feed quality, or productivity of individual cows or their herdmates change over time. For example, modeling herd-year-season contemporary groups in a time-dependent manner can account for the precise management conditions that will influence the risk of culling for a particular cow during each month of her life, and it is not necessary to assume that all cows calving in a given herd-year-season period will be subject to the same conditions throughout their entire lives.

Longevity data usually have a skewed distribution, and, in the absence of a suitable transformation, analysis of such data using methods that assume normality can lead to awkward results (Egger-Danner, 1993). Furthermore, the relationship between longevity and management or genetic factors is probably multiplicative rather than additive (Vukasinovic, 1999). Curiously, comparisons between survival analysis and linear model methodology based on actual data are lacking in the animal breeding literature.

The objectives of this study were to apply survival analysis methodology to the prediction of longevity PTA for US Jersey sires and to compare the accuracy of these predictions with those from the linear model approach that is currently used for routine national genetic evaluation of PL.

## MATERIALS AND METHODS

### Data

Culling and production data were provided by the USDA Animal Improvement Programs Laboratory. All cows were required to have valid sire identification and age at first calving between 18 and 42 mo. After editing, information data from 268,008 Jersey cows with first calving from 1981 to 2000 were available for the analysis.

Longevity was defined as the number of days from first calving until culling or censoring. Records from

cows that were sold for dairy purposes, cows residing in herds that discontinued milk recording, cows that were still alive after 5 completed lactations, and cows that were still alive at the time of analysis were considered censored. Cows that did not calve again within 6 mo after completing a "normal" lactation were considered dead and were treated as uncensored.

### Survival Analysis

The following Weibull proportional hazards (sire) model was used for analysis of length of PL:

$$h_{ijkl}(t) = h_0(t)\exp[P_i(t) + \beta A_j + h_{ysk}(t) + s_l] \quad [1]$$

where

$h_{ijkl}(t)$  = hazard function (instantaneous probability of culling) for a given cow at time  $t$ ;

$h_0(t)$  = Weibull baseline hazard function with scale parameter  $\tau$  and shape parameter  $\rho$ ;

$P_i(t)$  = time-dependent fixed parity-stage of lactation effect, assumed to be piecewise constant with change points at 0, 45, and 270 d of lactation 1, 2, 3, 4, or 5;

$A_j$  = time-independent fixed effect of age at first calving, treated as a continuous covariate with regression coefficient  $\beta$ ;

$h_{ysk}(t)$  = time-dependent random effect of herd-year-season, assumed to be independently distributed, following a log-gamma distribution with parameter  $\gamma$ , and assumed to be piecewise constant with change points at January 1, May 1, and September 1 of each year;

$s_l$  = time-independent random effect of sire of cow assumed to be distributed as multivariate normal with mean vector  $\underline{0}$  and covariance matrix  $\underline{A}\sigma_s^2$ , where  $\underline{A}$  is the additive relationship matrix between sires.

The Survival Kit version 3.12, a set of FORTRAN programs written by Ducrocq and Solkner (1998b), was used for the Weibull analysis. Details are given in Ducrocq (1994), and theoretical aspects are presented by Ducrocq and Casella (1996). An empirical Bayes approach was used to estimate the fixed parameters and to predict the random sire effects ( $s$ ).

The sire variance ( $\sigma_s^2$ ) was estimated as the mode of its marginal posterior density, which was approximated by Laplacian integration. A sire model was chosen for computational reasons, although the Survival Kit version 3.12 can accommodate an animal model as well. To avoid problems caused by herd-year-season classes

or sire progeny groups with few animals (and no uncensored failures), the parameters  $\gamma$ ,  $\sigma_s^2$  and  $\rho$  were estimated from a subset of data from 11 large herds and with a minimum of 20 uncensored daughters per sire. It is unknown whether such a restriction induces a sampling bias. To the extent that censoring is random over sires (an assumption of the procedure), we conjecture that such bias is negligible.

### Linear Model Analysis

Because one of our main goals was to compare survival analysis methodology with the linear model approach currently used for routine national genetic evaluation of PL in the US, we also estimated sire PTA for longevity using the following linear (animal) model:

$$y_{ijkl} = HYS_i + AC_j + a_k + e_{ijkl} \quad [2]$$

where

$y_{ijkl}$  = completed or projected months in milk at 84 mo of age for a given cow;

$HYS_i$  = fixed effect of herd-year-season of first calving;

$AC_j$  = fixed effect of age at first calving; and

$a_k$  = random additive genetic effect of animal  $k$  assumed to be distributed as multivariate normal with mean vector  $\underline{0}$  and covariance matrix  $\underline{A}$   $\sigma_a^2$ , where  $\underline{A}$  is the relationship matrix between animals and  $\sigma_a^2$  is the additive genetic variance.

Model [2] is similar to the linear model used by the USDA Animal Improvement Programs Laboratory, except that projected observations for censored animals were based only on age at first calving and current months in milk. (Information about current months dry or current milking status was lacking.) Genetic and residual variance components were estimated from the data via REML, and sire PTA were predicted by BLUP, conditionally, on the estimated variance components. Pearson correlations coefficients were obtained between PTA obtained in this study and PTA from the November 2003 USDA sire summary. For sires with at least 30 uncensored daughters, the correlation was 0.63, and for sires with at least 100 uncensored daughters and born before 1995, the correlation was 0.75.

### Model Comparisons

The data were split into 2 samples via random selection of herds; Subsets A and B contained records from 122,388 and 128,450 animals, respectively. Longevity PTA for the sires were subsequently predicted, conditionally, on estimates of the Weibull and dispersion

parameters obtained from all data, by applying Model [1] to each subset. The longevity PTA of each Jersey sire was expressed as the risk of culling of his daughters relative to the risk of culling for daughters of an average sire. Sire PTA from Model [1] for Subsets A and B will be denoted as  $RR_A$  and  $RR_B$ , respectively. Similarly, sire PTA were estimated in Subsets A and B using linear Model [2], and the resulting predictions will subsequently be referred to as  $PL_A$  and  $PL_B$ , respectively. Correlations between estimated sire PTA from each method (survival analysis or linear model) and each subset of the data (A or B) were calculated to assess the agreement (or lack thereof) between genetic predictions from each type of methodology.

Three different approaches were used for model validation. First, stayability observations (i.e., binary responses) indicating survival to second, third, fourth, or fifth lactation for cows in Subset A were regressed (separately) on  $RR_A$  or  $PL_A$  of their sires using logistic regression, such that each sire's PTA (from each model) could be converted into the probability that his daughters would survive to second, third, fourth, or fifth lactation. Next, the expected number of daughters of each sire that would survive to second, third, fourth, or fifth lactation in Subset B was calculated by multiplying the probability of survival to each lactation (from Subset A) by the total number of daughters in Subset B. The actual number of daughters in Subset B that survived to second, third, fourth, or fifth lactation in Subset B was subsequently determined for each sire, and the observed and expected number of "survivors" and "failures" were compared using the following  $\chi^2$  statistic:

$$\chi^2 = [(\text{observed survivors} - \text{expected survivors})^2 + (\text{observed failures} - \text{expected failures})^2].$$

These  $\chi^2$  statistics were summed across sires, and the model that produced the smallest sum was considered as the most accurate predictor of stayability. The reciprocal analysis was subsequently carried out by using survival probabilities computed from Subset B to predict daughters' actual survival rates in Subset A. Because some daughters of bulls born in recent years may have not yet had an opportunity to survive to third, fourth, or fifth lactation, we repeated the analysis using only those sires that were born in 1992 or earlier. Finally, a weighted analysis was conducted, in which each part of the  $\chi^2$  statistics for individual sires was weighted by the number of daughters of that sire.

In the second approach to cross-validation, the Kullback-Leibler (KL) criterion (Kullback, 1968) for measuring the distance between 2 distributions (one of them assumed to be the true distribution) was applied to the stayability data described previously. A rough interpre-

tation of the KL distance (Sorensen and Gianola, 2002) is the expected value (under the true distribution) of the log-odds ratio of the posterior probabilities of each of 2 models that are, a priori, equally likely.

Let  $f(\mathbf{y}|\theta_T)$  be the density of the “true” distribution, where  $\mathbf{y}$  is the data vector (the binary stayability variables in our case) and  $\theta_T$  is the parameter of the distribution. Let  $p(\mathbf{y}|\theta_A)$  be the density of an alternative distribution. The KL discrepancy between the 2 distributions is defined as

$$I(f, p) = \int \log \left[ \frac{f(\mathbf{y}|\theta_T)}{p(\mathbf{y}|\theta_A)} \right] f(\mathbf{y}|\theta_T) d\mathbf{y}.$$

If the data are discrete,  $f(\mathbf{y}|\theta_T)$  and  $p(\mathbf{y}|\theta_A)$  are probability distributions, and the KL discrepancy is

$$I(f, p) = \sum_{\mathbf{y}} \log \left[ \frac{f(\mathbf{y}|\theta_T)}{p(\mathbf{y}|\theta_A)} \right] f(\mathbf{y}|\theta_T).$$

In our study, we regressed (logistically) binary stayability observations of daughters in each subset on their sires’ relative risk (RR) ratios or PL PTA (from the survival and linear models, respectively). From these regressions, predicted probabilities of survival to second, third, fourth, and fifth lactations were obtained. Denote the predicted probability of survival (to the beginning of each lactation) for daughters of sire  $i$  as  $\hat{\theta}_{RR}^{(i)}$  and  $\hat{\theta}_{PL}^{(i)}$  for the survival and linear models, respectively. Next, consider the empirical distribution of the number of daughters that survived to a given lactation in the other subset of data as the “true” distribution. (This empirical distribution approaches the true distribution as the number of daughters per sire goes to infinity.) In other words, these empirical probabilities are calculated as

$$\hat{\theta}_T^{(i)} = \frac{n_{i,survive}}{n_i}$$

where  $n_i$  and  $n_{i,survive}$  are the total number of daughters of sire  $i$  and the number of daughters of sire  $i$  that survive to a particular lactation, respectively. We can regard the  $n_{i,survive}$  from different sires as independent random variables that follow a binomial ( $n_i, \hat{\theta}_T^{(i)}$ ) distribution. In practice, however, the empirical probabilities  $\hat{\theta}_T^{(i)}$  can be null if there are no surviving daughters. In this situation, the KL distance is not defined because the logarithm of 0 is not defined. Although it would be possible to exclude such cases for computational purposes, it is more intuitively appealing to “shrink” estimates based on limited information toward the population average because survival probabilities of null or

unity are not biologically plausible. If we assume that the binomial probabilities  $\hat{\theta}_T^{(i)}$  are distributed a priori across sires as Beta ( $a, b$ ), then the posterior distribution of  $\hat{\theta}_T^{(i)}$  is a Beta( $n_{i,survive} + a, n_{i,fail} + b$ ) distribution, where  $n_{i,survive}$  and  $n_{i,fail}$  represent the number of survivors and failures for sire  $i$ , respectively, and where the parameters  $a$  and  $b$  are estimated from the variation in relative frequencies between sires using a method of moments fit (Gelman et al., 2003); we refer to this as an empirical Bayes approach. We calculated the overall survival rate as

$$\hat{\theta} = \frac{\sum_{i=1}^M n_{i,stay}}{\sum_{i=1}^M n_i}$$

where  $M$  is the number of sires, and the between-sire variance in survival probabilities as

$$Var(\hat{\theta}_T^{(i)}) = \frac{\sum_{i=1}^M \theta_T^{(i)2}}{M} - \hat{\theta}^2 = V.$$

Following Gelman et al. (2003), we calculated

$$a + b = \frac{\hat{\theta}(1 - \hat{\theta})}{V} - 1$$

$$a = (a + b)\hat{\theta}$$

and

$$b = (a + b)(1 - \hat{\theta}).$$

The posterior mean of the probability of survival of daughters of sire  $i$  is

$$\tilde{\theta}_T^{(i)} = \frac{n_{i,stay} + a}{n_i + a + b}.$$

This was used as the true probability in the empirical distribution of the data, and the KL criteria was calculated as

$$I(T, A) = \sum_{i=1}^M n_i \left\{ \tilde{\theta}_T^{(i)} \log \frac{\tilde{\theta}_T^{(i)}}{\tilde{\theta}_A^{(i)}} + (1 - \tilde{\theta}_T^{(i)}) \log \frac{1 - \tilde{\theta}_T^{(i)}}{1 - \tilde{\theta}_A^{(i)}} \right\}$$

where  $\tilde{\theta}_A^{(i)}$  is set equal to  $\tilde{\theta}_{RR}^{(i)}$  or  $\tilde{\theta}_{PL}^{(i)}$  when calculating the KL distance (alternative distribution) for the survival analysis or linear model, respectively.

The third approach was quite similar to the logistic regression and  $\chi^2$  statistic described previously. How-

ever, instead of measuring stayability to the beginning of second, third, fourth, or fifth lactation for daughters of a particular sire, we considered stayability to 36, 48, 60, 72, and 84 mo of life. Only those daughters that had an opportunity period of 36, 48, 60, 72, or 84 mo, respectively, were included in calculation of the  $\chi^2$  statistic.

## RESULTS AND DISCUSSION

For the 268,008 Jersey cows included in this study, 45% of the records were censored. The mean failure time was 807 d after first calving, and the mean censoring time was 954 d after first calving. Parameter estimates were 8.2 ( $\gamma$ ), 0.94 ( $\rho$ ), and 0.021 ( $\sigma_s^2$ ). Heritability on the logarithmic scale was approximated using the equation of Ducrocq and Casella (1996):

$$h_{\log}^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \Psi^{(1)}(\gamma) + \frac{\pi^2}{6}}$$

where  $\Psi^{(1)}(\gamma)$  = tri-gamma function evaluated at  $\gamma$ , and  $\frac{\pi^2}{6}$  is the variance of an extreme value distribution. This yielded an estimated heritability on the log scale of 4.7%. Heritability on the original scale was approximated using a Taylor series expansion of  $\log h^2$  around its mean, as in Ducrocq and Casella (1996):

$$h^2 = \frac{4\sigma_s^2}{\left[e\left(\frac{1}{\rho}\right)v\right]^2\left(\sigma_s^2 + \Psi^{(1)}(\gamma) + \frac{\pi^2}{6}\right)} + \frac{h_{\log}^2}{\left[e\left(\frac{1}{\rho}\right)v\right]^2}$$

where  $v = \Psi(\gamma) - \log(\gamma) - \text{Euler's constant}$ , and  $\Psi(\gamma)$  = di-gamma function evaluated at  $\gamma$ . This yielded a heritability estimate on the original scale of 18.0%. By comparison, the estimated heritability of PL using the linear Model [2] was 7.0%. The approximation to heritability on the original scale typically gives estimates that are larger than those obtained from linear models. This suggests that either environmental influences are modeled more precisely using this survival analysis methodology or, perhaps, that the approximation given previously is inadequate (Korsgaard et al., 2002).

Survival analysis methodology allows expressing a sire's genetic merit in several ways, including the RR of culling, the expected percentage of daughters still alive after a given number of lactations, or the expected length of PL. In the latter case, estimates correspond to points on a given sire's survival curve,  $S(t)$ . For example, the median survival time occurs when  $S(t) = 0.50$ . In this study, we used RR of culling to express genetic

**Table 1.** Number of cows in each subset of the data by birth year of their sires.

Birth year of sire	Cows	
	Subset A	Subset B
	(no.)	
1984	11,664	11,543
1985	31,863	31,788
1986	10,624	12,747
1987	9089	9680
1988	16,270	18,884
1989	10,164	11,388
1990	6656	6632
1991	5023	5449
1992	3727	3972
1993	3870	4330
1994	3283	3071
1995	956	837
Total	113,189	120,321

merit for survival. For example, daughters of a sire with a RR of culling of 1.2 are expected, on average, to have a 20% greater chance of being culled at any given time than daughters of an average sire (with RR of culling = 1.0). For the Jersey sires considered herein, risk ratios for individual sires ranged from approximately 0.7 (low risk of culling) to 1.3 (high risk of culling).

The number of cows in each data set, by birth year of their sires, is shown in Table 1, and correlations in predicted genetic merit between methods (survival or linear model) and data sets (A or B) are shown in Table 2. These correlations were 0.73 for  $RR_A$  and  $RR_B$ , 0.76 for  $PL_A$  and  $PL_B$ , -0.55 for  $PL_A$  and  $RR_A$ , -0.60 for  $PL_B$  and  $RR_B$ , -0.46 for  $PL_A$  and  $RR_B$ , and -0.35 for  $PL_B$  and  $RR_A$ . These results indicate a stronger correlation within methodologies than within data sets (with different methodologies).

Table 3 shows the actual number of cows that survived to second, third, fourth, and fifth lactation in each data set. On average, 64 to 66% of the cows survived to second lactation, 38 to 41% survived to third lactation, 21 to 23% survived to fourth lactation, and 10 to 11% survived to fifth lactation.

Table 4 shows the sums of  $\chi^2$  statistics from the observed vs. expected number of survivors to second,

**Table 2.** Correlations between predicted genetic merit of Jersey sires for longevity obtained from random Subsets A and B using survival analysis ( $RR_A$  and  $RR_B$ ) or a linear model ( $PL_A$  and  $PL_B$ ).

	$RR_B$	$PL_A$	$PL_B$
$RR_A$			
$RR_B$	0.73		
$PL_A$	-0.55	-0.46	
$PL_B$	-0.35	-0.60	0.76

**Table 3.** Number and percentage of Jersey cows that survived to second, third, fourth, or fifth lactation in each of the random Subsets A and B.

Lactation	All Sires					
	Subset A			Subset B		
	Culled	Survived	(%)	Culled	Survived	(%)
2	41,237	71,952	0.64	43,255	77,066	0.64
3	70,124	43,065	0.38	74,217	46,104	0.38
4	89,576	23,613	0.21	94,904	25,417	0.21
5	101,787	11,402	0.10	108,099	12,222	0.10
	Sires born prior to 1992					
2	34,667	66,686	0.66	36,801	71,310	0.66
3	59,900	41,453	0.41	63,688	44,423	0.41
4	77,931	23,422	0.23	82,859	25,252	0.23
5	89,951	11,402	0.11	95,889	12,222	0.11

third, fourth, and fifth lactation. When using estimated survival probabilities from Subset A to predict the survival rate of daughters to the beginning of each lactation in Subset B, the RR ratios from the survival analysis (sire model) were superior to the productive life PTA from the linear (animal) model analysis in all cases. When using estimated probabilities from Subset B to predict daughters' survival in Subset A, the linear model gave better predictions of survival to second and third lactation, but the survival analysis results were superior predictors of survival to fourth and fifth lactation. However, some daughters of sires born in recent

years may have not yet had an opportunity to survive to third, fourth, or fifth lactation. Therefore, the analysis was repeated using only sires born prior to 1992. When this restriction was applied, survival analysis was a more accurate predictor of stayability to each lactation in both Subsets A and B.

Because some dairy bulls are used much more extensively than others as sires of replacement animals, we also conducted a weighted analysis in which the  $\chi^2$  statistic was weighted by the number of progeny for each sire. Results (Table 5) were quite similar to the

**Table 4.** Sum of  $\chi^2$  statistics over sires based on comparing the predicted vs. observed number of survivors to each lactation in each subset of the data (using probabilities of survival derived from the other subset). The method providing the most accurate predictions is shown in bold in each case.

	All sires (n = 2019)		
	Lactation	Linear model	Survival analysis
Prediction of survival in Subset B from sire PTA calculated in Subset A	2	3,763,785	<b>3,624,011</b>
	3	5,713,067	<b>5,509,714</b>
	4	4,440,178	<b>3,692,692</b>
	5	1,986,108	<b>1,411,709</b>
Sires born prior to 1992 (n = 1445)			
	2	3,707,714	<b>3,119,807</b>
	3	5,872,924	<b>4,739,221</b>
	4	4,639,546	<b>3,196,978</b>
	5	2,067,985	<b>1,278,696</b>
All sires (n = 1991)			
Prediction of survival in Subset A from sire PTA calculated in Subset B	2	<b>2,505,513</b>	3,318,580
	3	<b>4,089,887</b>	4,600,435
	4	3,240,735	<b>2,728,493</b>
	5	1,449,913	<b>1,095,660</b>
Sires born prior to 1992 (n = 1433)			
	2	6,143,712	<b>3,965,734</b>
	3	8,570,678	<b>7,559,236</b>
	4	3,598,130	<b>3,304,571</b>
	5	1,999,708	<b>1,794,798</b>

**Table 5.** Weighted sum of  $\chi^2$  statistics over sires (weighted by the number of daughters per sire) based on comparing the predicted vs. observed number of survivors to each lactation in each subset of the data (using probabilities of survival derived from the other subset). The method providing the most accurate predictions is shown in bold in each case.

	All sires (n = 2019)		
	Lactation	Linear model	Survival analysis
Prediction of survival in Subset B from sire PTA calculated in Subset A	2	103,535	<b>94,337</b>
	3	165,829	<b>152,795</b>
	4	148,914	<b>105,326</b>
	5	74,397	<b>41,374</b>
Sires born prior to 1992 (n = 1445)			
	2	118,167	<b>77,625</b>
	3	201,147	<b>126,346</b>
	4	182,895	<b>90,829</b>
	5	89,866	<b>38,894</b>
All sires (n = 1991)			
Prediction of survival in Subset A from sire PTA calculated in Subset B	2	<b>42,498</b>	77,366
	3	<b>94,559</b>	107,259
	4	103,026	<b>67,132</b>
	5	51,929	<b>28,610</b>
Sires born prior to 1992 (n = 1433)			
	2	<b>49,206</b>	73,878
	3	112,732	<b>94,261</b>
	4	124,466	<b>55,953</b>
	5	62,562	<b>25,217</b>

**Table 6.** Estimated Kullback-Leibler differences (using a Beta distribution) based on comparing the predicted vs. observed number of survivors to each lactation in each subset of the data (using probabilities of survival derived from the other subset). The method providing the most accurate predictions is shown in bold in each case.

All sires (n = 2019)			
	Lactation	Linear model	Survival analysis
Prediction of survival in Subset B from sire PTA calculated in Subset A	2	<b>5077</b>	5338
	3	<b>7642</b>	8042
	4	<b>7005</b>	7242
	5	<b>5189</b>	5316
	- Sires born prior to 1992 (n = 1445) -		
	2	6374	<b>5370</b>
	3	8915	<b>7658</b>
	4	7113	<b>7082</b>
	5	<b>8195</b>	8771
All sires (n = 1991)			
Prediction of survival in Subset A from sire PTA calculated in Subset B	2	<b>9341</b>	9595
	3	<b>10,617</b>	11,844
	4	<b>8397</b>	9979
	5	<b>8969</b>	9656
	- Sires born prior to 1992 (n = 1433) -		
	2	6007	<b>5418</b>
	3	<b>7409</b>	7446
	4	<b>5827</b>	6277
	5	7566	<b>7417</b>

unweighted analysis; survival analysis tended to be a better predictor of stayability than the corresponding linear model analysis. As expected, weighting by the number of daughters in the Jersey breed resulted in a very large contribution to the  $\chi^2$  from a few sires (with many daughters). However, this does not compromise the usefulness of the results because these sires are important to the Jersey breed.

Results of the KL distance comparison are shown in Table 6. When all sires were included, the probabilities from the linear model analysis agreed more closely with the empirical distribution of stayability in both Subsets A and B than the probabilities from the survival analysis. However, restricting the comparison to sires born before 1992 indicated that survival analysis was a better predictor of survival to second, third, and fourth lactation in Subset B and to first and fourth lactation in Subset A.

Table 7 shows results of the third approach, in which logistic regression and  $\chi^2$  statistics were used to compare models predicting stayability of daughters to 36, 48, 60, 72, and 84 mo of life among daughters that had opportunity to survive that long. Survival analysis was a better predictor of stayability than linear model analysis in all cases in Subset A. In Subset B, survival analysis was a superior predictor of stayability to 48, 60, and 72 mo of life, but the linear model was a better predictor of stayability to 36 or 84 mo of life.

**Table 7.** Sum of  $\chi^2$  statistics over sires based on comparing the predicted vs. observed number of survivors to 36, 48, 60, 72, and 84 mo of age among daughters that had an opportunity to survive that long in each subset of the data (using probabilities of survival derived from the other subset). The method providing the most accurate predictions is shown in bold in each case.

	Age (mo)	Linear model	Survival analysis
Prediction of survival in Subset B from sire PTA calculated in Subset A	36	194,952	<b>159,821</b>
	48	167,943	<b>120,657</b>
	60	158,681	<b>113,763</b>
	72	125,021	<b>87,057</b>
	84	49,782	<b>39,311</b>
Prediction of survival in Subset A from sire PTA calculated in Subset B	36	<b>201,861</b>	233,689
	48	195,144	<b>160,325</b>
	60	176,211	<b>124,469</b>
	72	136,271	<b>118,828</b>
	84	<b>42,661</b>	44,616

## CONCLUSIONS

Survival analysis methodology is theoretically superior to linear model approaches for the analysis of longevity data because it allows proper treatment of censored observations, inclusion of time-dependent covariates, and a skewed or non-normal distribution of survival times. These statistical properties alone justify consideration of this methodology for routine national genetic evaluation of dairy cow longevity.

Results from the present study suggest that survival analysis can yield higher heritability estimates than linear models, which may lead to more rapid genetic progress, provided that such heritability approximations are meaningful (Korsgaard et al., 2002). Further, differences occurred in sire rankings between the 2 methods considered in this study, so a change in methodology would alter the usage level of certain sires in commercial dairy herds.

None of the 3 methods ( $\chi^2$  analysis of predicted stayability to a certain lactation, KL distance comparison, and  $\chi^2$  analysis of predicted stayability to a certain age) provided a clear-cut answer regarding the superiority or inferiority of survival analysis methodology. However, predictions derived from estimates of sires' genetic merit for longevity from the survival analysis tended to be more closely related to actual stayability observations and culling times in independent samples of the data. It should be noted that these comparisons favored the linear model, a priori, because more pedigree information was used in the linear (animal) model than in the survival (sire) model. When combined with the aforementioned theoretical advantages, these results tend to support the adoption of survival analysis methodology for routine genetic evaluation of longevity in dairy cattle. Although survival analysis methodology is computationally demanding, particularly when using

an animal model, advances in computing speed and power should allow its implementation on a national scale. Furthermore, any advantages in reliability or stability of longevity evaluations for dairy sires attributable to implementation of this methodology would be obtained with no additional data collection costs.

### ACKNOWLEDGMENTS

Financial support of the National Association of Animal Breeders (Columbia, MO), the American Jersey Cattle Association (Reynoldsburg, OH), and The Babcock Institute for International Dairy Research and Development (Madison, WI) is greatly appreciated. Technical assistance was provided by Vincent Ducrocq, and data were provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD).

### REFERENCES

- Ducrocq, V. 1987. An analysis of length of productive life in dairy cattle. Ph.D. Diss., Cornell Univ., Ithaca, NY.
- Ducrocq, V. 1994. Statistical analysis of length of productive life for dairy cows of the Normande breed. *J. Dairy Sci.* 77:855–866.
- Ducrocq, V., and G. Casella. 1996. A Bayesian analysis of mixed survival models. *Genet. Sel. Evol.* 28:505–529.
- Ducrocq, V., and J. Solkner. 1998a. Implementation of a routine breeding value evaluation for longevity of dairy cows using survival analysis techniques. *Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia* 27:447–448.
- Ducrocq, V., and J. Solkner. 1998b. The Survival Kit—V3.0; A Package for Large Analyses of Survival Data. *Vil. Proc. 6th World Congr. Genet. Appl. Livest. Prod., Armidale, Australia* 22:51–52.
- Egger-Danner, C. 1993. Zuchtwertschätzung für Merkmale der Langlebigkeit beim Rind mit Methoden der Lebensdaueranalyse. Ph.D. Diss., Univ. für Bodenkultur, Vienna, Austria.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. *Bayesian Data Analysis*, 2nd ed. CRC Press, Boca Raton, FL.
- Korsgaard, I. R., A. H. Andersen, and J. Jensen. 2002. Prediction error variance and expected response to selection, when selection is based on the best predictor—For Gaussian and threshold characters, traits following a Poisson mixed model and survival traits. *Genet. Sel. Evol.* 34:307–333.
- Kullback, S. 1968. *Information Theory and Statistics*. Dover Publications, New York, NY.
- Smith, S. P., and R. L. Quaas. 1984. Productive lifespan of bull progeny groups: Failure time analysis. *J. Dairy Sci.* 67:2999–3007.
- Sorensen, D., and D. Gianola. 2002. *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York, NY.
- VanRaden, P. M., and E. J. H. Klaaskate. 1993. Genetic evaluation of length of productive life including predicted longevity of live cows. *J. Dairy Sci.* 76:2758–2764.
- Vukasinovic, N. 1999. Application of survival analysis in breeding for longevity. *Interbull Bull.* No. 21, Uppsala, Sweden.