The creation and evaluation of a model predicting the probability of conception in seasonal-calving, pasture-based dairy cows

Caroline Fenlon,*1,2 Luke O’Grady,†1 Michael L. Doherty,† John Dunnon,* Laurence Shalloo,‡ and Stephen T. Butler‡

*School of Computer Science, and
†School of Veterinary Medicine, University College Dublin, Belfield, Dublin 4, Ireland
‡Animal and Grassland Research and Innovation Centre, Teagasc, Moorepark, Fermoy, County Cork, Ireland

ABSTRACT

Reproductive performance in pasture-based production systems has a fundamentally important effect on economic efficiency. The individual factors affecting the probability of submission and conception are multifaceted and have been extensively researched. The present study analyzed some of these factors in relation to service-level probability of conception in seasonal-calving pasture-based dairy cows to develop a predictive model of conception. Data relating to 2,966 services from 737 cows on 2 research farms were used for model development and data from 9 commercial dairy farms were used for model testing, comprising 4,212 services from 1,471 cows. The data spanned a 15-yr period and originated from seasonal-calving pasture-based dairy herds in Ireland. The calving season for the study herds extended from January to June, with peak calving in February and March. A base mixed-effects logistic regression model was created using a stepwise model-building strategy and incorporated parity, days in milk, interservice interval, calving difficulty, and predicted transmitting abilities for calving interval and milk production traits. To attempt to further improve the predictive capability of the model, the addition of effects that were not statistically significant was considered, resulting in a final model composed of the base model with the inclusion of BCS at service. The models’ predictions were evaluated using discrimination to measure their ability to correctly classify positive and negative cases. Precision, recall, F-score, and area under the receiver operating characteristic curve (AUC) were calculated. Calibration tests measured the accuracy of the predicted probabilities. These included tests of overall goodness-of-fit, bias, and calibration error. Both models performed better than using the population average probability of conception. Neither of the models showed high levels of discrimination (base model AUC 0.61, final model AUC 0.62), possibly because of the narrow central range of conception rates in the study herds. The final model was found to reliably predict the probability of conception without bias when evaluated against the full external data set, with a mean absolute calibration error of 2.4%. The chosen model could be used to support a farmer’s decision-making and in stochastic simulation of fertility in seasonal-calving pasture-based dairy cows.

Key words: dairy cow, fertility, milk production, logistic regression model, external evaluation

INTRODUCTION

Herd fertility is a key driver of economic efficiency in pasture-based production systems (Shalloo et al., 2014). The main goal of reproductive management in pasture-based production systems is to achieve a calving pattern that most closely aligns the herd energy requirements during lactation with pasture growth during the grazing season. This strategy reduces the need for additional, more costly feeds to meet energy demands (Butler, 2014). Accurate knowledge of the probability of conception to an individual service would allow confident precision management of breeding to optimize herd fertility performance and profitability.

The individual factors affecting both submission and conception have been extensively researched. Although interrelated, these factors include genetics, milk production, nutrition and energy balance, calving events, reproductive tract disorders, infectious diseases, other diseases such as lameness and mastitis, sire fertility, AI technique, and farm management (Walsh et al., 2011). Genetics has a strong influence on the potential fertility performance of a cow and can modulate the response to environmental influences such as nutrition and disease. In 2001, the Irish national breeding program changed from being predominantly focused on milk production traits to a more holistic economically weighted multi-trait index called the Economic Breeding Index (EBI; Veerkamp et al., 2002). The EBI fertility subindex

Received August 5, 2016.
Accepted March 18, 2017.
1These authors contributed equally to this study.
2Corresponding author: caroline.fenlon@ucdconnect.ie
comprises 2 traits: calving interval and survival. Favorable genetic merit for fertility traits requires negative PTA for calving interval and positive PTA for survival (Berry et al., 2007). Excessive or prolonged negative energy balance is associated with greater loss of body-weight and BCS, decreased milk protein concentration and increased milk fat to protein ratios (Roche et al., 2009). The reproductive consequences of excessive negative energy balance include reduced odds of establishment and maintenance of pregnancy (Lucy, 2001; Butler, 2014).

Accurate predictive models can be useful as tools for stand-alone decision support or as parts of larger simulation models. Previous prediction models of dairy cow conception to individual services have been created using various machine learning techniques. These include logistic regression, naïve Bayes, Bayesian networks, decision trees, random forests, support vector machines, and partial least squares regression (Caraviello et al., 2006; Shahinfar et al., 2014; Hempstalk et al., 2015; Rutten et al., 2016). The models developed by Caraviello et al. (2006) and Rutten et al. (2016) aimed to predict first service outcomes only, whereas the models of Hempstalk et al. (2015) and Shahinfar et al. (2014) predicted all service outcomes. Only one previous study has attempted to predict conception in seasonal-calving pasture-based production systems (Hempstalk et al., 2015), while the other studies used data from year-round breeding systems. Important variables commonly reported in predictive models include average farm conception rate, year and season, and cow-level variables such as parity, DIM, service number, BCS, milk yield, and health events. Hempstalk et al. (2015) and Shahinfar et al. (2014) also included genetic values for milk yield and fertility in their models.

Validation of prediction outcomes differ among studies, using bootstrapping (Rutten et al., 2016) or 5-fold (Shahinfar et al., 2014) or 10-fold (Caraviello et al., 2006; Hempstalk et al., 2015) cross-validation. Hempstalk et al. (2015) also used test data held back from the model building process to evaluate predictions using 2 different approaches, either using year or herd to segment their data. None of the cited studies used independent data sets external to the model building process to validate the predictions from their models.

The accuracy of these prediction models is most often reported using area under the curve (AUC) and the proportion of correctly classified instances. From the studies cited above, Random Forest models have best prediction performance with moderate AUC values, ranging from 0.64 (Hempstalk et al., 2015) to 0.76 (Shahinfar et al., 2014). Hempstalk et al. (2015) also found logistic regression performed favorably (AUC 0.622). Rutten et al. (2016) used logistic regression only, but no overall measure of model accuracy was reported. Instead, evaluations were performed by comparing the mean predicted probability of conception to the mean conception rate within categorical variable groups (e.g., BCS).

The objective of the present study was to create and externally evaluate a predictive model of the probability of conception to any given service, primarily for use in a whole-farm decision-support tool of dairy cow fertility.

**MATERIALS AND METHODS**

**Data**

Two independent data sets from Irish seasonal-calving pasture-based dairy herds were available for analysis, one for model training and the other for testing. For model training, data were sourced from the centralized database at Teagasc’s Animal and Grassland Research and Innovation Centre, Moorepark, County Cork, Ireland. The animals included in the data set were from the Curtins and Ballydague spring-calving research herds, which were representative of Irish grass-based farming systems with a diverse range of cow genetics (Horan et al., 2005). The records consisted of 6,008 AI services for 1,076 lactating cows (3,047 lactations) from 2001 to 2009. Additional data available included individual animal calving records, hormonal treatments, genetic information, live weight and BCS, daily milk production records (including yield and composition), and results of reproductive tract examinations. An external data set was used for testing of model predictions. Records were available from 9 herds of Holstein-Friesian dairy cows involved in a herd health and fertility consultancy program operated by the School of Veterinary Medicine, University College Dublin (Somers et al., 2015) from 2008 to 2015 and consisted of 11,262 AI service records (6,259 lactations of 2,752 animals). Data available for these herds included individual animal calving records, genetic information, BCS, milk production records (including yield and composition), and results of reproductive tract examinations.

Herds in both data sets operated seasonal-calving pasture-based production systems, with 67.5% of the cows calving in February or March. Details regarding average annual herd size, calving pattern, parity, BCS, milk production, and fertility performance are detailed in Table 1.

**Data Cleaning and Calculations**

All data manipulation and analyses were carried out using the R statistical programming language (R Core Team, 2015).
### Table 1. Descriptive statistics at herd level for model training and testing data sets [mean (SD) where appropriate]; data span 2001 to 2009 and 2009 to 2014 for training and testing data sets, respectively

<table>
<thead>
<tr>
<th>Variable</th>
<th>Training herd</th>
<th>Testing herd</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Herd size</td>
<td>184.80 (36.47)</td>
<td>112.33 (21.11)</td>
</tr>
<tr>
<td>Parity</td>
<td>2.62 (1.60)</td>
<td>2.23 (1.28)</td>
</tr>
<tr>
<td>Service conception rate (%)</td>
<td>50.03 (50.01)</td>
<td>41.85 (49.34)</td>
</tr>
<tr>
<td>Services per lactation</td>
<td>1.71 (0.94)</td>
<td>1.87 (0.98)</td>
</tr>
<tr>
<td>In-calf within 42 d (%)</td>
<td>61.79 (6.01)</td>
<td>54.43 (9.72)</td>
</tr>
<tr>
<td>Service BCS</td>
<td>3.34 (0.36)</td>
<td>3.26 (0.37)</td>
</tr>
<tr>
<td>Calving BCS</td>
<td>2.88 (0.31)</td>
<td>2.93 (0.29)</td>
</tr>
<tr>
<td>Standardized service milk yield (kg)</td>
<td>28.31 (6.77)</td>
<td>25.41 (6.07)</td>
</tr>
<tr>
<td>Service milk fat (%)</td>
<td>4.07 (0.50)</td>
<td>4.10 (0.69)</td>
</tr>
<tr>
<td>Service milk protein (%)</td>
<td>3.35 (0.25)</td>
<td>3.40 (0.31)</td>
</tr>
<tr>
<td>305-d yield</td>
<td>5,766.88 (997.27)</td>
<td>5,328.23 (978.65)</td>
</tr>
<tr>
<td>Calvings (%)</td>
<td>25.1 (43.3)</td>
<td>20.1 (43.3)</td>
</tr>
<tr>
<td>January</td>
<td>18.3 (43.3)</td>
<td>10.0 (43.3)</td>
</tr>
<tr>
<td>February</td>
<td>27.0 (43.3)</td>
<td>37.4 (43.3)</td>
</tr>
<tr>
<td>March</td>
<td>21.7 (43.3)</td>
<td>34.3 (43.3)</td>
</tr>
<tr>
<td>April</td>
<td>24.2 (43.3)</td>
<td>16.1 (43.3)</td>
</tr>
<tr>
<td>May</td>
<td>6.8 (43.3)</td>
<td>2.2 (43.3)</td>
</tr>
<tr>
<td>June</td>
<td>1.9 (43.3)</td>
<td>0.6 (43.3)</td>
</tr>
</tbody>
</table>
**Service Dates and Outcomes.** Service data included animal tag number, lactation number, calving and service dates, service number, service sire, and pregnancy outcome. All services were performed by AI. The outcome in the original data was defined within the database as being successful if the last service within an individual lactation had a subsequent calving date, regardless of duration. All other services were defined within the database as negative. Gestation lengths were calculated for all positive services, and interservice intervals were calculated within each cow’s lactation. During the data-cleaning process, gestation length and interservice intervals were recalculated immediately after the removal of any service event. Initially, histograms of service data were visually inspected to look for biologically implausible values (e.g., served and conceiving within a few days of calving) and extreme outlying interservice intervals. From the initial inspection, data relating to services were edited as follows. Any service occurring before 30 DIM (training n = 108, testing n = 100) was removed. Any service after 170 DIM was deleted to remove any carryover cows from previous calving seasons (training n = 80, testing n = 2,332). Any services occurring within less than 3 d of a previous service were treated as a single estrus event, and only the first service was retained (training n = 248, testing n = 131). If a cow had more than one service removed because of short repeat intervals during a particular lactation, then all services for that cow’s lactation were removed (training n = 85).

Using farm treatment records, services following hormonal treatments (and the lactation during which they occurred) were identified. This included lactations during which doses of PGF$_{2\alpha}$, after the first service (n = 373 services, 133 lactations), GnRH (n = 24 services, 10 lactations), and intravaginal progesterone-releasing device (CIDR; n = 262 services, 143 lactations) were administered.

To validate the outcome of each service in the original data set, 2 separate methods were used to define a successful conception. Where ultrasound pregnancy diagnosis records were available, a service was defined as successful if a diagnosis of a viable fetus was made after 30 d and the cow had no subsequent service or ultrasound diagnosis of embryonic loss made within 60 d of the initial service. In the absence of any ultrasound examination, a service was defined as successful if it was the last service for the lactation and the cow had a subsequent calving event within 282 ± 15 d or a recorded abortion event after 60 d. The final service during an individual cow’s lactation was discarded if an ultrasound pregnancy diagnosis or subsequent calving date was not available to define an outcome. For these cows, services before the final service were retained and defined as unsuccessful.

If the calving date corresponded more closely to an earlier service date, later services were removed and the service previously marked as negative but corresponding to a calving date was changed to successful. If the calculated gestation length from a final service was defined as too long (≥298 d), it was considered that it must have been followed by another unrecorded service (and conception). These final services were removed as outlined above (training n = 73, testing n = 349).

Explanatory variables were grouped into the following categories.

**Individual Animal Characteristics, Treatments, and Previous Fertility.** Explanatory variables included the continuous variables age (days) and DIM at service, and the categorical variables parity (1, 2, 3, 4, ≥5), year/month/day of service, number of services (1, 2, 3, 4, ≥5), and the number of days since previous service (interservice interval) grouped into the following bins: ≤17, 18–24, 25–35, 36–48, and ≥49 d. Hormonal treatments before service were included as binary variables.

As an indicator of previous fertility performance, the number of days open in the previous lactation was calculated as the length of time in days to conception from the mating start date of the previous year, or previous calving date if calving occurred after the mating start date in the previous year. For first lactation cows, this value was set to 0.

**Calving Events.** The following calving event details were available for each lactation: calving month and day of the year; calving difficulty score using an ordinal scale (1 = normal calving; 2 = some assistance; 3 = considerable difficulty; 4 = veterinary assistance); calf birth weight in kilograms; and binary variables indicating multiple calves, abortion, and stillbirth. Data on retained fetal membranes were not recorded.

**Milk Yield and Composition.** Daily milk production was measured an average of 256 times per animal lactation. Fat and protein concentrations were recorded on average 71 times per lactation (am and pm sampling) in the training data. Milk yield and composition were measured on average 6.3 times per lactation in the testing data.

The milk yield (kg/d) data were standardized to 40 g/kg fat and 31 g/kg protein (MY$_{\text{standardized}}$) using the following calculation (Faverdin et al., 2012):

$$
MY_{\text{standardized}} = \frac{MY_{\text{ACT}}}{0.44 + 0.0055(FY - 40) + 0.0033(PY - 31)}
$$
where $\text{MY}_{\text{ACT}}$ = recorded yield in kilograms, $\text{FY}$ = recorded milk fat yield in grams, and $\text{PY}$ = recorded milk protein yield in grams.

For days when milk fat and protein contents were not measured, missing values were estimated using splines (by cow/lactation, average training $n = 68$) created with the “smooth.spline” function from the R “stats” package (R Core Team, 2015). The predictions of the splines were evaluated using the root mean squared error function ($\text{RMSE}$; Hyndman and Koehler, 2006).

Missing service-day milk records were estimated by taking the average standardized milk yields of the closest surrounding records if yields were recorded within 3 d of the service ($n = 805$). Additional missing standardized milk yield values were predicted using the MilkBot model (Ehrlich, 2012), which fits yield curves for each individual cow adjusted by lactation (training $n = 198$). Because fewer milk recordings were taken in the testing herds, MilkBot was used to estimate all service-day milk yields.

**BW and BCS.** Body weight was measured on average 36 times during a cow’s lactation in the training herds but was not measured in the testing herds. The mean number of BCS measurements taken per lactation was 6 (SD 7.6) in the training herds and 4 (SD 2.6) in the testing herds.

Where a BW record on the day of calving was not available, the closest recording $\leq 10$ d post calving was assigned as the calving weight. Where BW was not recorded at service, the closest recording within $\pm 10$ d of the service date was assigned to the service. If a BW record could not be assigned then the weight was considered missing for the corresponding service.

Where calving BCS data were not available on the day of calving, the closest BCS to calving within 20 d before calving was used instead. If no BCS data were available within this range, the highest score from 1 mo before calving to 10 d post calving was used. If no BCS score record was present within these time periods, then a calving BCS score was deemed missing (training $n = 1,846$, testing $n = 3,502$).

Because the number of records close to each service was low, BCS at service was instead estimated as a continuous variable using splines based on all of the cow’s measurements during the lactation. Body condition score variables were considered in the form of continuous and discrete values with continuous scores at service rounded to the nearest 0.25 unit of BCS, and values $< 2.75$ and $> 4.00$, which had few records, were grouped.

For measurements of both cow weight and BCS, average daily changes between recordings were calculated and visual assessment was used to define cutoff values for the removal of outliers. With the exception of the day of calving, weight measurements with an absolute average daily difference of greater than 50 kg between consecutive recordings were removed. Body condition scores with an absolute average daily difference of greater than 0.15 units between consecutive recordings were removed. Average daily rate of weight change on the day of AI service was calculated as the average daily change in body weight, using the weight measurements taken closest to the date of the service within 28 d.

The BW and BCS variables were calculated: at calving and at each service date; nadir BW and BCS; current and maximum loss to date at each service; rate of BW change $\pm 30$ and $\pm 7$ d around service; DIM at nadir; and days since nadir.

**Breed and Genetics.** The PTA values for all traits used in the calculation of the Irish national breeding program EBI values were sourced in July 2014 from the Irish Cattle Breeders Federation (ICBF) national database for all cows in the data set. These included the overall EBI and the subindexes for milk production, fertility ($\text{FERT}_\text{SI}$), calving, beef, maintenance ($\text{MAINT}_\text{SI}$), and health ($\text{HEALTH}_\text{SI}$) as well as PTA values for progeny carcass weight, progeny carcass conformation, progeny carcass fat score ($\text{CFAT}$), cull cow carcass weight ($\text{CULL}$), cow weight ($\text{LWT}$), direct calving difficulty, gestation length, calf mortality, maternal calving difficulty, milk yield in kilograms, milk fat percentage ($\text{FAT}$ %), milk protein percentage, milk fat yield, milk protein yield, SCS, locomotion score, calving interval ($\text{CIV}$), and survival ($\text{SUR}$). A description of how the PTAs for individual traits contribute to each of the subindexes is available from the ICBF website (ICBF, 2010). Details on the individual calculation of PTA and economic weightings have been described by Berry et al. (2007). Records on breed composition for each individual cow (i.e., percentage primary and secondary breeds) were also available as defined in the ICBF database, with Holstein and Friesian being the predominant primary breeds. Other breeds included Montbéliarde, Normande, and Norwegian Red.

**Statistical Methods and Data Analysis**

**Modeling Strategy.** The predictive logistic regression model was developed in 2 phases to allow an evaluation of the wide range of data available, while maintaining a high level of interpretability of the model. Initially, univariate analysis was used to screen variables, followed by stepwise multivariate mixed-effects model building (Phase 1). This was followed by testing the predictive ability of the phase 1 base model against the
inclusion of all combinations of variables selected from the univariate analysis to identify the best prediction model (Phase 2). The data set originating from the research herds was chosen for training the predictive model because, despite having fewer services and herds, it contained a more comprehensive range of factors to evaluate as part of the model-building process.

**Phase 1: Descriptive and Univariate Analysis.** The mean and standard deviation were calculated for each of the continuous variables available in the data set, and counts of services were included for categorical variables. For continuous variables, the distribution of values was checked for normality and analyzed for outliers (biologically impossible or unlikely values). Given the large number of variables available for analysis (n = 58) and therefore the >1600 potential predictive models (before the inclusion of interactions), univariate analysis was used to screen candidate variables for inclusion in the multivariate model. The univariate relationship between each of the explanatory variables with the probability of conception was estimated using binary logistic regression ["glm" function (R Core Team, 2015)]. Because of the nonlinear pattern of their relationships with probability of conception, DIM and service BCS variables were transformed. Log DIM was a more suitable fit than the untransformed DIM, and a second-order polynomial (BCS + BCS²) had the best fit for service BCS. These relationships were confirmed by testing their linearity in the logit probability of conception.

All variables were compared for correlations using Pearson correlation coefficient for comparisons of continuous variables and ANOVA or chi-squared tests for continuous vs. categorical and categorical vs. categorical comparisons, respectively.

**Phase 1: Multivariate Analysis and Model Building.** Following the univariate analysis, a multivariate mixed-effects binary logistic regression model was constructed. Variables with a P-value ≤ 0.20 in the univariate analysis were considered for inclusion as fixed effects in the multivariate model. A base model was created using the most significant variables (in univariate analysis) from each of the different explanatory variable categories. The model was then refined by removing variables not significant at a P-value ≤ 0.20 (using the R “drop1” function) and adding variables or interactions of clear biological importance that had a significance level of P ≤ 0.05 (“add1” function). The testing of the variables was repeated until no more were available to improve the model. Once no more individual variables could be included or excluded from the model, all 2-way interactions between the variables in the fixed effects model were tested. All factors potentially affecting conception were tested. All factors potentially affecting conception in an unquantified way were considered for inclusion as random effects, using the “glmer” function of the lme4 R package (Bates et al., 2015). These included cow, service sire, hormonal treatments, month or year of service, and herd. Cow, sire, and year within herd were found to be the most useful and were retained in the final models.

Highly correlated pairs of variables (correlation coefficient >0.7) were identified and the most significant variable from the univariate analysis was initially tested in the multivariate model with the correlated variable excluded. The correlated variable was tested for inclusion at the final stage of model building when all excluded variables were re-entered into the final model to evaluate any effects of multicollinearity. Moderately correlated variables (correlation coefficient ± 0.3 to 0.7) within each group were eligible for inclusion in the multivariate model, and the effects of multicollinearity were assessed as each variable was entered or removed. To evaluate possible effects of multicollinearity in the multivariate model, variance inflation factors were calculated using the “vif” function from the car package (Fox and Weisberg, 2011).

**Phase 2: Identifying Best Prediction Model.** To attempt to further improve the predictive capability of the base model developed in phase 1, all potential combinations of the variables considered during the multivariate model building process of phase 1, but not found to be significant, were tested by adding them to the base model. To select the best predictive model, the prediction evaluation procedures outlined below were performed and compared with the base model developed in phase 1. To evaluate model predictions, the probability of conception and the resultant binary outcome were predicted for each service in the test data set using the base and final models.

**Model Evaluation.** Measures of discrimination and calibration were used to evaluate the model predictions. Discrimination evaluates the ability of a model to distinguish between positive and negative insemination outcomes, with a threshold of 50% used to transform the predicted probabilities into binary insemination outcomes. Calculations were all based on the tabulation of predicted outcomes as the number of true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) events. Precision [TP/(TP + FP)] identifies how many of the services that were predicted as successful actually occurred (Olson and Delen, 2008). Recall [TP/(TP + FN)] calculates the proportion of the actual successful services that were identified as such by a model (Olson and Delen, 2008).

Precision and recall were combined to calculate the F-measure, the harmonic mean of the 2 scores (van Rijsbergen, 1979):
Matthews correlation coefficient was calculated to measure if each model’s discrimination was better than a random selection (Matthews, 1975):

$$\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

For each model, a receiver operating characteristic (ROC) curve was created by plotting the trade-off between the FP and TP rate as the discrimination threshold is altered (“ROCR” R package; Sing et al., 2005). The area under the ROC curve (AUC) indicates the probability that the model will predict a higher probability for a randomly chosen service that resulted in conception than for a randomly chosen service that resulted in failure to conceive.

Calibration techniques test the accuracy of the predicted probabilities; that is, they determine if the observed frequency of actual conceptions is similar to the predicted probability within groups of service records (Hosmer et al., 2013). Calibration can also identify bias: over- or under-estimation of the predicted probabilities. The Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980) was used to evaluate the overall goodness-of-fit of the models on the testing data (using the “ResourceSelection” R package; Johnson et al., 2006). Calibration plots were created using 25 equidistant probability groups, with F-distribution confidence intervals (“PresenceAbsence” R package; Freeman and Moisen, 2008). Deviance values were calculated as the difference between the actual service outcome (coded as 1 or 0) and the predicted probability of conception and arranged into groups of equal size sorted by predicted probability (“arm” R package; Gelman and Su, 2015). The average group deviances were plotted against the predicted probability. The absolute values of the deviances were averaged to find the mean absolute calibration error (MACE), which indicates the average error of the predicted probabilities.

RESULTS

Estimation of Missing Values

The RMSE of the known values used to build each spline for predicting missing data was calculated. The overall test-day RMSE of the MilkBot predictions, within the range of DIM involved in the analysis (30–169 DIM), was 2.7 kg/d for the training set and 4.3 kg/d for the testing set. Average RMSE by DIM became progressively lower by later lactation. The average test-day RMSE of the splines for daily milk fat and protein concentrations in the training data set were 8.6 and 1.2 g/kg, respectively. The average RMSE of the splines used to predict service BCS was 0.067 units for the training data and 0.05 units for the testing data.

Phase 1: Univariate Analysis

After cleaning and imputation of missing values, 5,176 services (2,934 lactations from 1,059 animals) were retained for inclusion in the model building.

In the univariate analysis, increasing log DIM, BCS at calving, nadir BCS, second-order polynomial of service BCS, percentage Friesian breed, EBI, FERT_SI, MAINT_SI, HEALTH_SI, and PTA for FAT %, CFAT, SUR, and SCS and decreasing values for CIV PTA were significantly associated ($P < 0.05$) with an increase in the probability of conception. Increasing age, parity, service number, calving difficulty score, calf birth weight, number of days open in the previous lactation, percentage Holstein breed, CULL PTA, and PTA variables related to calving difficulty and milk production were associated with decreased probability of conception. All categories of interservice intervals, with the exception of the 18- to 24-d group, were associated with a reduced probability of conception. Twinning and abortions were also associated with a reduced probability of conception. Additional variables considered for inclusion in the multivariate modeling process ($P < 0.2$) showed tendencies for decreased probability of conception when cows had greater LWT PTA, weight at calving or service, standardized milk yield, milk fat to protein ratio at service, and maximum BCS loss since calving. Decreased probability of conception was also observed in cows with lesser PROT % PTA and cows that had been given any hormonal treatment.

Phase 1: Multivariate Mixed Model

The training data set was missing 3,162 service-day BCS, 275 PTA values, and 1,141 calving difficulties, leaving 2,723 records for the model-building process. The base multivariate model (Table 2), based on step-wise modeling techniques, contained parity, log DIM, interservice interval, calving difficulty score, CIV PTA, milk kg PTA, and the interaction between CIV PTA and log DIM. The standard deviations of the random effects, with mean 0 by default, were 0.21 (cow), 0.2 (service sire), and 0.25 (year within herd).

Phase 2: Best Multivariate Mixed Prediction Model

After evaluation of the multiple possible models, a final prediction model was identified, consisting of the
Comparison of Base and Final Models

After cleaning, 4,212 service records (1,471 animals from 9 herds) remained for testing. Missing data primarily consisted of BCS (1,856 services missing).

**Discrimination Analysis**

Both models performed similarly in discrimination metrics with the final model marginally outperforming the base model in all metrics (Table 4). The inspection of ROC curves showed that the 2 models performed very similarly with no probability superior to the standard 50% cutoff for prediction.

**Calibration Analysis**

The calibration plots for the models are shown in Figure 1. Both of the models passed the Hosmer-Lemeshow test of goodness-of-fit using $P > 0.05$ ($P = 0.08$ for the base model, $P = 0.57$ for the final model). Eighty percent of the base model’s deviance values were positive, indicating some bias in the predicted probabilities. The MACE of the final model was 2.4%, 0.3 percentage points lower than the base model (Table 4).

**Herd-Level Evaluation**

When individually applied to the herds within the testing data set, predictions from the final model for 7 of the 9 testing herds (mean $n = 361$) were found to have no significant bias (calibration lines with slope 1 and intercept 0) and had MACE ranging from 4.3 to 7.3%, with a mean of 6.3%, and AUC ranging from 0.59 to 0.65 with a mean of 0.61. Herd 7 had showed some bias, with a calibration line showing a slope of 1 but an intercept higher than 0 due to a higher farm effect on conception probability. The MACE and AUC for herd 7 ($n = 745$) was 11.0% and 0.63, respectively. Herd 2 ($n = 928$) also showed bias, with a calibration line slope lower than 1. The MACE and AUC for herd 2 were 8.2% and 0.57, respectively.

**DISCUSSION**

We have analyzed, in combination, a wide array of detailed phenotypic and genetic factors related to fertility. The effects on fertility of many of the risk fac-
tors have been previously studied in seasonal-calving pasture-based systems and other farming systems. Our study provides novel insights into the factors affecting conception throughout lactation and the interactions between them. Our study is also novel in the evaluation of model predictions. In predictive modeling studies closely comparable to our own (Shahinfar et al., 2014; Hempstalk et al., 2015), models were evaluated using data related to the data set used for model training, either via bootstrapping or data splitting methods. To our knowledge, our study is the first to create predictive models of conception to any given service in seasonal-calving pasture-based dairy cows and evaluate them using an independent external data set sourced from nonresearch herds. We are also the first to employ calibration methods for the model evaluation process, thus gaining deeper insights into predictive performance.

The final variables in our model are in general agreement with model findings of Hempstalk et al. (2015) and Shahinfar et al. (2014), with DIM, parity, genetics, and BCS all strongly influencing prediction outcomes. The AUC value from the present study is the same as that obtained by Hempstalk et al. (2015) using logistic regression and is similar to those obtained by Shahinfar et al. (2014) using Bayesian or decision tree methods. In contrast to the other previous studies, no effect of seasonality, either season of calving or breeding, featured as important in our analysis. This finding may be due to the narrow breeding window in our data and the inclusion of year as a random effect, as typically the seasonal effects, such as weather, tend to differ more from year to year and do not necessarily follow a predictable month by month pattern. Hempstalk et al. (2015) also suggested that differing seasonal effects may exist in pasture-based production systems than in those modeled by Shahinfar et al. (2014). Both Shahinfar et al. (2014) and Hempstalk et al. (2015) made use of a herd-level prior probability of conception, in the current and previous seasons, respectively. In both cases, the prior had a very influential effect. In the study by

<table>
<thead>
<tr>
<th>Variable</th>
<th>Logit estimate</th>
<th>SE</th>
<th>Odds ratio (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−4.47</td>
<td>0.96</td>
<td>1.00</td>
<td>0.003</td>
</tr>
<tr>
<td>Parity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>−0.08</td>
<td>0.12</td>
<td>0.92 (0.74–1.16)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>−0.06</td>
<td>0.14</td>
<td>0.94 (0.72–1.23)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>−0.45</td>
<td>0.16</td>
<td>0.64 (0.47–0.88)</td>
<td></td>
</tr>
<tr>
<td>≥5</td>
<td>−0.58</td>
<td>0.18</td>
<td>0.56 (0.39–0.80)</td>
<td></td>
</tr>
<tr>
<td>CIV PTA</td>
<td>−0.50</td>
<td>0.20</td>
<td>0.61 (0.41–0.89)</td>
<td></td>
</tr>
<tr>
<td>Log DIM</td>
<td>1.03</td>
<td>0.21</td>
<td>2.81 (1.84–4.27)</td>
<td></td>
</tr>
<tr>
<td>Interservice interval (d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st service</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3–17</td>
<td>−0.58</td>
<td>0.21</td>
<td>0.56 (0.37–0.85)</td>
<td></td>
</tr>
<tr>
<td>18–24</td>
<td>−0.12</td>
<td>0.12</td>
<td>0.88 (0.69–1.13)</td>
<td></td>
</tr>
<tr>
<td>25–35</td>
<td>−0.79</td>
<td>0.19</td>
<td>0.46 (0.31–0.67)</td>
<td></td>
</tr>
<tr>
<td>36–48</td>
<td>−0.54</td>
<td>0.19</td>
<td>0.58 (0.40–0.84)</td>
<td></td>
</tr>
<tr>
<td>≥49</td>
<td>−1.00</td>
<td>0.31</td>
<td>0.37 (0.20–0.70)</td>
<td></td>
</tr>
<tr>
<td>Milk kg PTA/100</td>
<td>−0.05</td>
<td>0.03</td>
<td>0.95 (0.90–1.01)</td>
<td>0.096</td>
</tr>
<tr>
<td>Calving difficulty</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal calving</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some assistance</td>
<td>−0.18</td>
<td>0.11</td>
<td>0.83 (0.67–1.03)</td>
<td></td>
</tr>
<tr>
<td>CD/VA</td>
<td>−0.36</td>
<td>0.15</td>
<td>0.70 (0.52–0.94)</td>
<td></td>
</tr>
<tr>
<td>CIV PTA × log DIM</td>
<td>0.10</td>
<td>0.04</td>
<td>1.10 (1.01–1.20)</td>
<td>0.020</td>
</tr>
<tr>
<td>BCS at service</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First order</td>
<td>3.35</td>
<td>2.43</td>
<td>28.35 (2.24–3.416.89)</td>
<td></td>
</tr>
<tr>
<td>Second order</td>
<td>−1.13</td>
<td>2.21</td>
<td>0.32 (0.00–2.489)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Base model</th>
<th>Final model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (%)</td>
<td>56.56</td>
<td>56.84</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>48.12</td>
<td>48.87</td>
</tr>
<tr>
<td>AUC</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>F-measure (%)</td>
<td>52.00</td>
<td>52.55</td>
</tr>
<tr>
<td>Matthews correlation coefficient</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>Mean absolute calibration error (%)</td>
<td>2.73</td>
<td>2.42</td>
</tr>
<tr>
<td>Hosmer-Lemeshow P-value</td>
<td>0.08</td>
<td>0.57</td>
</tr>
</tbody>
</table>

1Milk kg = milk production PTA (kg); CIV = calving interval PTA (d); CD/VA = considerable difficulty/veterinary assistance.
Figure 1. Calibration plots of predictions on external test data for 2 mixed logistic regression models, base model (top) and final model (bottom). Point size reflects relative number of services in the group. The shaded area indicates the 95% CI.
groups of services with similar attributes, is centered
the probability of conception, both overall and within
income is to be expected in such data, however, where
ing between negative and positive outcomes. This out-
based on the standard 50% threshold for discriminat-
in predicting the outcome of an individual service us-
predictive models of conception discriminate very well
poor prediction (Cox and Snell, 1989). This analysis,
tests is that they cannot identify more specific cases of
rate of the group and no bias was evident.
not significantly differ from the actual mean conception
but the mean predicted probabilities within groups did
produced undetected bias, because although their model
herd prior as stated above. They may have also intro-
undetected bias, because although their model
performed well in animals with higher predicted prob-
abilities, it performed no better than random guessing
when applied to cows with lower predicted probabil-
ities. Unfortunately, the results reported by Rutten et
al. (2016) preclude a direct comparison of the models,
but the mean predicted probabilities within groups did
not significantly differ from the actual mean conception
rate of the group and no bias was evident.
To allow the assessment of both accuracy and po-
tential bias we used discrimination tests, which are
commonly applied in data mining, and the less com-
mon calibration methods. Measures of discrimination
are used to test the capability of the models to accu-
rately split the data into positive and negative service
outcomes. The disadvantage of overall goodness-of-fit
tests is that they cannot identify more specific cases of
poor prediction (Cox and Snell, 1989). This analysis,
along with previous research, demonstrates that no
predictive models of conception discriminate very well
in predicting the outcome of an individual service us-
ing a particular combination of animal characteristics
based on the standard 50% threshold for discriminat-
ing between negative and positive outcomes. This out-
come is to be expected in such data, however, where
the probability of conception, both overall and within
groups of services with similar attributes, is centered
near 50% and will very rarely be extremely high or low.
In these circumstances, a more thorough investigation
of predictive capabilities within groups should be used
in conjunction with the discrimination tests. Although
the Hosmer-Lemeshow test found no significant differ-
ces between the predicted probabilities and actual
conception rates in either phases of our model building,
a more stringent P-value requirement (e.g., \( P = 0.1 \))
would result in the base model (with \( P = 0.08 \)) failing
this goodness-of-fit test. The final model showed the
greatest accuracy and showed no significant evidence of
bias in predictions. Consequently, our prediction model
is more generalizable and suited for use in stochastic
modeling and decision-support tools, where it is more
useful to know the probability of success of a service
with a particular set of characteristics rather than a
prediction of an absolute outcome. For example, rank-
ing of services that are more likely to be successful will
help the farmer when allocating expensive and scarce
resources such as sexed semen and high genetic merit
sires.
Our final predictive model on the probability of
conception contained the effects of parity, stage of lac-
tation, interservice intervals, BCS, and genetic traits
for fertility and milk production. Direct comparison of
odds ratios with other studies is difficult because of
methodological differences regarding what constitutes
a successful conception or pregnancy, but our findings
are in general agreement with reported findings across
a range of farming systems (Gröhn and Rajala-Schul-
tz, 2000; Xu and Burton, 2003; Hagiya et al., 2013;
Shahinfar et al., 2014). In our model, the probability
of conception was favorably associated with superior
genetic traits for fertility (CIV PTA) and unfavorably
associated with superior genetic traits for milk produc-
tion (milk kg PTA). This association was particularly
influential for CIV PTA and showed an interaction with
DIM. Lower CIV PTA not only resulted in a higher
influential for CIV PTA and showed an interaction with
DIM. Lower CIV PTA not only resulted in a higher
effect was most evident in early lactation. For example,
comparing a cow with a CIV PTA of −4 d with a cow
with a CIV PTA of −3 d (all else being equal), the
probability of conception is predicted to be 3 percent-
ages between the predicted probabilities and actual
ences between the predicted probabilities and actual

ability of conception. These include greater circulating concentrations of IGF-1 throughout lactation, more rapid postpartum uterine recovery, earlier postpartum resumption of cyclicity, increased pre-ovulatory follicle diameter, larger corpus luteum volume, and greater luteal phase circulating progesterone concentrations.

The main areas where our results differ from some previous studies relate to the associations between probability of conception and BCS and milk production. Surprisingly, no measures associated with energy balance (BCS, milk protein, or milk fat to protein ratio) that were significant in the univariate analysis were retained as statistically significant effects in the base model. However, BCS at service did feature in the final predictive model, suggesting that BCS has an important effect and that absolute BCS at the time of service is more strongly associated with reproductive performance than BCS at calving or the magnitude of BCS loss after calving. The lack of effect of calving BCS or BCS loss disagrees with the findings of Buckley et al. (2003) and Roche et al. (2007). However, the analysis used in the present study included all services and not just the first service post-calving. The relative importance of BCS loss after calving compared with BCS at service will likely decline the further into lactation that a cow progresses. As mentioned previously, genetic merit for fertility traits has been associated with the physiological regulation of energy balance and reproduction during lactation. The inclusion of CIV PTA in the models may have accounted for some of the variability of the severity of negative energy balance. It is also possible that an inadequate range in BCS at calving or BCS loss was present in the data set for detecting a statistically significant difference.

In our study, phenotypic milk yield (unlike milk kg PTA) was not statistically significant in the model. Although including phenotypic milk yield did not improve the model predictions in phase 2 of the model creation process, the coefficient did show a tendency ($P < 0.1$) toward improved probability of conception with increasing yield. Although genetic selection for milk production has been linked to adverse effects on genetic merit for fertility, evidence for a direct effect of phenotypic milk yield or milk solids production on reproduction is less clear, with some studies showing reduced (Westwood et al., 2002; Evans et al., 2006), improved (Buckley et al., 2003; Xu and Burton, 2003; Cummins et al., 2012a), or no association (Patton et al., 2007) between fertility and milk production. Our findings are consistent with previous research, in which accounting for unmeasured herd effects resulted in no detectable association between milk yield and reproductive performance (Bello et al., 2012, 2013). The observed association of improved reproductive performance with increased milk yield has been hypothesized to be due to improved herd management and health (Buckley et al., 2003; Bello et al., 2012).

Service number was identified in many of the previously cited studies as a significant factor affecting conception. Service number is likely confounded by DIM and the prior probability of conception in the individual cow, rather than an actual effect of previous insemination. In our study, the inclusion of DIM, genomics, and previous service intervals effects have likely accounted for the effects that may be associated with service number.

The accuracy of our model may be limited because of the nature of the data available for this study. A key assumption was made that the service events were performed at a true estrus event (i.e., the probability of conception was greater than 0). This assumption was obviously valid when conception was successful but otherwise cannot be truly defined in our study. We attempted to avoid including service records from cows that were not in estrus via the removal of lactations with multiple abnormal repeat intervals and the removal of services with very short interservice intervals. The inclusion of interservice intervals and their significant association with reduced probability of conception in biologically abnormal interservice intervals in the models is likely to account for some of the effects of service error, while acknowledging that some cows may have genuine estrous cycles that fall outside the 18- to 24-d range (Macmillan and Watson, 1971; Remnant et al., 2015). A potential limitation of the scope of the present study in modeling conception was the absence of data on clinical disease (e.g., uterine health, production diseases, and infectious diseases) because these have all been reported to have significant unfavorable effects on reproductive performance (Fourichon et al., 1999; McDougall, 2001; McDougall et al., 2007; LeBlanc, 2008; Walsh et al., 2011). Unmeasured effects at the herd level, such as management, may also limit the accuracy of our model predictions. The model contains several significant random effects at both the animal and herd level to account for herd-to-herd variation. Although this inclusion allows prediction of the average cow/herd/year, these factors may differ substantially in some herds. Overall, our model accurately generalized across the testing herds, but some evidence of variation was present. This variation can be partially addressed if historical on-farm data are available for recalibration of the model coefficients.

Although the present study included a limited number of herds, given that the cow type and production systems of the farms used to generate and evaluate these models are typical of those throughout Ireland, our model findings are likely generalizable to most sea-
sonal pasture-based dairy herds in Ireland. All of the factors included in the model are within the capabilities of all farmers to record. Because the final model contained variables consistent with the literature from other seasonal-calving dairy-producing countries, with similar reproductive management and performance (Dairy Australia., 2011; Xu and Burton, 2003), it is probable that this model would be transferable internationally.

CONCLUSIONS

The risk factors contained in the final model to predict the probability of conception in seasonal-calving pasture-based dairy cows include both genetic and phenotypic variables and interactions between them. Key findings from our analysis of risk factors include the interaction of CIV PTA with DIM, lending further support to previous work showing differing adaptations to lactation with differing CIV PTA, and the significance of milk yield kg PTA in the absence of any effect of phenotypic milk yield on conception. These findings raise important questions for future research regarding the underlying nature of these effects and interactions. Our study also successfully demonstrated the benefits of using calibration methods to evaluate the predictive ability of logistic regression models, and we suggest that these methods be routinely used in addition to the traditional discrimination methods to evaluate future models. Last, the final model, which is based on easily recordable and obtainable data, provided an accurate estimate of the probability of conception. This model will be valuable for both prediction and simulation of dairy cow fertility in seasonal-calving pasture-based dairy production systems.

ACKNOWLEDGMENTS

This research was funded by Dairy Research Ireland. The authors thank Anne Geoghegan of Teagasc Moorepark for assistance in gathering the data used in this study. We also thank the farmers involved in the University College Dublin School of Veterinary Medicine’s herd fertility consultancy program for their cooperation and data collection.

REFERENCES


