# Challenges when analysing a large data set to identify determinants of dairy cow reproductive performance

**Morton, J.M.[1] and Anderson, G.A.[2]**

[1]School of Veterinary Science, The University of Queensland, Queensland Australia
[2]Department of Veterinary Science, The University of Melbourne, 250 Princes Highway, Werribee, Victoria Australia

## Abstract

When studying determinants of dairy cow reproductive performance, it is desirable to fit insemination-level variables when modelling lactation-level outcomes such as pregnancy by week 6 after mating start date. Insemination-level variables may be confounding effect estimates for some herd- and lactation-level variables. In addition, insemination-level variables may explain additional variance in outcome.

This can be achieved by fitting lactation-level categorical variables for each insemination-level variable but this approach effectively incorporates some of the 'pattern of inseminations' for each cow into the model. The pattern of inseminations is described by the number of inseminations and the day of the mating period in which each insemination is performed, and is a strong determinant of pregnancy by week 6 after mating start date. Estimates of the total effect for risk factors for pregnancy by week 6 of the mating period that are mediated at least partly through pattern of inseminations may be underestimated using this approach. In addition, estimates of the proportion of variance in days to conception that is explained by this model would overestimate the proportion of variance explained by risk factors that can be manipulated by herd managers.

An alternative approach is discussed, using a two-level multi-level survival (frailty) model (lactation within herd) with insemination-level variables fitted as time-varying covariates. This would allow simultaneous modelling of herd-, lactation- and insemination-level variables, removing any confounding of herd- and lactation-level variables by insemination-level variables. However further work is required to ascertain whether this approach avoids both underestimation of the total effect for risk factors for pregnancy that are mediated at least partly through pattern of inseminations and overestimation of the proportion of variance in days to conception explained by risk factors that can be manipulated by herd managers.

## Introduction

Reproductive performance was described and risk factors and population attributable risks for reproductive performance were studied in a large multi-centred prospective observational study, conducted in 168 commercial dairy herds from throughout Australia's eastern states.

The objectives of the study were to:

- to describe reproductive performance in study herds,
- to identify risk factors associated with reproductive performance and to estimate the strength of association for these factors, and
- to estimate population attributable fractions and risks for factors associated with pregnancy by week 6 after mating start date.

A further objective was to assess the proportion of variation between herds that was explained by the final statistical models. This objective was of industry importance. If a large proportion of variation between herds is explained by the factors included in statistical models, this would suggest that additional independent unmeasured risk factors were relatively unimportant across the study

herds. This paper describes a methodological challenge to achieving these objectives and suggests a possible approach to address this challenge.

## Materials and methods

### *Herd and lactation selection*

One hundred and seventy herds were enrolled and 168 herds (124 seasonal calving herds and 44 year-round calving herds) completed the study. In these 168 herds, 34,631 lactations commenced with a calving during the study period. After exclusions and losses to follow-up, 29,462 lactations were retained for analyses. Conception dates were estimated by manual rectal pregnancy diagnosis. Most cows diagnosed pregnant were estimated to have conceived between 5 and 15 weeks previously.

### *Outcome definition*

Primary outcomes were based on time from each cow's mating start date to conception. In year-round calving herds, mating start date was defined for each cow as calving date plus herd voluntary waiting period. Within each seasonal calving herd, each cow was allocated the same calendar mating start date. Two binary outcome variables - pregnancy by week 6 and non-pregnancy by week 21 after mating start date, were chosen as primary lactation-level outcomes of interest after consideration of distributions of proportions pregnant by week of mating, economic aspects, ease of communication, appropriateness for both seasonal and year-round calving production systems and, in seasonal calving herds, durations of artificial insemination and total mating periods. In year-round calving herds, primary measures can also be based on time from calving. Pregnancy by day 100 and non-pregnancy by day 200 after calving are alternative primary lactation-level outcomes of interest. However, as these measures are inappropriate for seasonal calving herds, risk factors for these outcomes were not modelled. Only results of models of pregnancy by week 6 after mating start date are discussed further in this paper.

### *Modelling approach*

Associations between pregnancy by week 6 after mating start date and multiple exposure variables (putative risk factors) were assessed using multivariable multilevel logistic models. Following screening of lactation-level variables, a lactation-level model was built using ordinary logistic regression in SAS version 6.12 (SAS Institute, Cary, North Carolina). Herd-level factors were then screened using a multilevel model with the final lactation-level model fitted. The final multilevel model was then built using MLwiN 1.10.0006 (Multilevel Models Project, Institute of Education, University of London, UK) with lactations (level 1) clustered within herds (level 2) using the logit link and binomial error distribution. The multilevel model was estimated with the restricted iterative generalised least squares (RIGLS) procedure using second-order penalised quasi-likelihood (PQL) estimates. The level 1 variance was defined as $\bullet^2/3$ or 3.29 (Snijders and Bosker 1999).

## Variances 'explained' by the final model

The (lactation-level) outcome variance was 0.24 and only a low proportion of this variance was 'explained' by the final model. This model explained 18.4% of the outcome variance; 2.3% was unexplained at the herd-level and 79.3% was unexplained at the lactation-level. However the random (ie herd-level) intercept variance reduced substantially from 0.306 (se 0.038) in the null model to 0.096 (se 0.014) in the final model, indicating that the final model explained a substantial proportion of the variation in the outcome between herds.

## Possible approaches to include insemination-level variables

The final model that caused this substantial reduction in the random (ie herd-level) intercept variance did not include insemination-level variables, yet such variables have been associated with probability of conception to first insemination in previous studies as well as in other components of the current study. These variables include artificial insemination technician type (professional or farm technician) and characteristics of the semen such as bull identity. Insemination-level variables may be confounding effect estimates for some herd- and lactation-level variables. In addition, it seems likely that selected insemination-level variables could explain an additional proportion of the outcome variance and, more importantly, result in further reductions in the random (ie herd-level) intercept variance. Thus it is desirable to include insemination-level variables in these models.

Because the outcome is at the lactation- rather than insemination-level, including insemination-level variables raises challenges. One possible analytical approach is to include insemination-level variables as lactation-level variables. For example, the variable representing technician type at first service can be fitted as a dichotomous variable. However, this data is only available for cows that receive at least one insemination in the first 6 weeks of the mating period. For the current study, this would require exclusion of approximately 15% of cows. Alternatively, a dummy variable can be fitted to account for non-submitted cows.

Both of these approaches only use data from first services but a substantial proportion of cows receive two or more inseminations in the first 6 weeks of the mating period. To incorporate data from first and second inseminations, a more complex lactation-level categorical variable could be constructed (Table 1) and associated dummy variables fitted.

**Table 1 Levels in a lactation-level categorical variable to describe artificial insemination technician type (professional or farm technician) at first and second inseminations in the first 6 weeks of the mating period**

| | |
|---|---|
| No insemination in the first 6 weeks of the mating period | |
| First insemination by professional technician | No second insemination |
| First insemination by professional technician | Second insemination by professional technician |
| First insemination by professional technician | Second insemination by farm technician |
| First insemination by farm technician | No second insemination |
| First insemination by farm technician | Second insemination by professional technician |
| First insemination by farm technician | Second insemination by farm technician |

Application of this latter approach to variables with numerous levels would result in an associated lactation-level categorical variable with a large number of levels. For example, semen from a wide range of bulls was used in the study, with semen from 30 different bulls accounting for only half of all inseminations. The number of levels of the lactation-level categorical variable created to summarise an insemination-level variable with n levels can be as high as $n^2 + n + 1$.

A more important limitation of all of these approaches is that they effectively incorporate some of the 'pattern of inseminations' for each lactation into the model. The pattern of inseminations for each lactation is described by the number of inseminations and the day of the mating period in which each insemination is performed. Number of inseminations in the first 6 weeks of the mating period is a strong determinant of pregnancy by week 6 after mating start date (Table 2). The reduction in proportion of cows pregnant by week 6 after mating start date in cows with more inseminations reflects the interaction between the number of inseminations and period of mating when each insemination occurred.

Given the futility of inseminating non-ovulating cows and the need to minimise false positive inseminations in pregnant cows, pattern of inseminations is not directly determined by herd

managers. Rather, pattern of inseminations is an intervening variable between factors that can be manipulated by herd managers and occurrence of pregnancy by week 6 after mating start date. By effectively incorporating the pattern of inseminations for each cow into the model, an increased proportion of the outcome variance would probably be explained. However, it is more relevant from an industry perspective to estimate the proportion of the outcome variance explained by risk factors that can be manipulated by herd managers. Known determinants of pattern of inseminations are unlikely to explain all of the variance in pattern of inseminations. So, by effectively including the pattern of inseminations in the model, the proportion of the outcome variance explained by risk factors that can be manipulated by herd managers could be substantially overestimated. In addition, many risk factors for pregnancy by week 6 of the mating period are mediated at least partly through pattern of inseminations. By incorporating pattern of inseminations into the model, the strength of association for the 'total' effects of these risk factors would be underestimated (possibly substantially).

**Table 2 Association between number of inseminations by week 6 after mating start date and risk of pregnancy by week 6 after mating start date**

| No. inseminations by week 6 after mating start date | No. lactations | % pregnant by week 6 of mating period |
|---|---|---|
| 0 | 3,029 | 1.4% |
| 1 | 12,383 | 67.1% |
| 2 | 4,517 | 57.1% |
| 3 | 398 | 51.5% |
| 4-6 | 29 | 44.8% |
| Total | 20,356 | 54.8% |

## An alternative approach

An alternative approach is to fit a two-level multi-level survival (frailty) model (lactation within herd) with insemination-level variables fitted as time-varying covariates. This would allow simultaneous modelling of herd-, lactation- and insemination-level variables, removing any confounding of herd- and lactation-level variables by insemination-level variables. However this approach also effectively fits the pattern of inseminations for each lactation. Accordingly, estimates of the total effect for risk factors for pregnancy that are mediated at least partly through pattern of inseminations may still be underestimated. In addition, estimates of the proportion of variance in days to conception that is explained by this model may still overestimate the proportion of variance explained by risk factors that can be manipulated by herd managers.

## Reference

Snijders TAB and Bosker RJ (1999) *Multilevel analysis. An introduction to basic and advanced multilevel modelling.*, Sage Publications, London