VALIDATION OF CALVING EASE EBVS EXAMINING THE IMPACT OF GENETIC GROUPS AND SINGLE-STEP ON PREDICTIVE ABILITY

P.M. Gurman, L. Li, M.G. Jeyaruban, D.J. Johnston, C.J. Girard, and A.A. Swan

Animal Genetics Breeding Unit*, University of New England, Armidale, NSW, 2351 Australia

SUMMARY

Calving difficulty scores recorded in beef cattle are challenging to analyse due to low frequency of difficult births and the scored nature of the trait, requiring analysis in a threshold model, typically in combination with two linear traits, birth weight and gestation length. Previous software to calculate estimated breeding values (EBVs) for calving ease was not able to include single-step methods or fit genetic groups in models of analysis. In this study, we examined the value of including genetic groups and genomic information via single-step genomic BLUP (ssGBLUP) in the TransTasman Angus Cattle Evaluation (TACE) BREEDPLAN and Hereford BREEDPLAN analyses, by forward-validation in genotyped animals. The greatest improvements in accuracy were observed when including genomic information, with increases of 0.169 and 0.106 in the Angus and Hereford analyses respectively. Adding genetic groups to models had no impact on accuracy, but increased the bias of CE EBVs in ssGBLUP analyses for both breeds.

INTRODUCTION

Traits that are measured as scores are often difficult to analyse, especially if the distribution of the scores is skewed. A linear model can be used in some cases if the scores approximate normality, but a threshold model is typically used to address the imbalance in measurement between categories (Hoeschele *et al.* 1995; Gilmour *et al.* 1998). Mixed-model threshold analyses add extra complexity to solving for fixed and random effects due to the requirement of estimating both the threshold values and the weights to apply to each categorical phenotype.

Calving difficulty scores in BREEDPLAN analyses are characterised by low frequencies of difficult births. Analyses of this trait are performed using a categorical threshold model with birth weight and gestation length included as correlated linear traits to improve prediction for overall calving ease. Since 2017, BREEDPLAN analyses for most traits have been transitioning to ssGBLUP. In November 2019, a ssGBLUP implementation for calving ease was developed in new software for the BREEDPLAN component of the TransTasman Angus Cattle Evaluation (TACE, herein Angus), including genetic groups.

As part of the process of developing these enhancements, the utility of genetic groups came into question. The addition of genetic groups was observed to substantially increase convergence times of the model in the Angus evaluation, and when applied to Hereford BREEDPLAN, resulted in changes in EBVs that were difficult to interpret.

This paper examines the predictive ability of threshold model calving ease EBVs in Angus and Hereford BREEDPLAN with the inclusion of genetic groups and single-step using forward validation procedures.

MATERIALS AND METHODS

Calving ease (CE) data from the March 2022 Angus and May 2021 Hereford BREEDPLAN calving ease analyses after cleaning were used in this study. CE is scored as 1: no assistance required, 2: easy pull, 3: hard pull. Genetic parameters used for these models were adapted from Jeyaruban *et al.* (2015), with the genetic group variance assumed to be equal to the genetic variance. Genetic

* A joint venture between NSW Department of Primary Industries and University of New England

groups were fitted as routinely constructed in BREEDPLAN based on year window, breed, and country for the main analysis traits, with 20 groups for Angus and 16 for Hereford. There groups were included to improve prediction for animals with missing pedigree. Each data set was split into two groups, "training" and "validation", based on year of birth. The training set included animals born before 2019, while the validation set included animals with phenotypes born from 2019 onwards. BLUP analyses were performed in a factorial design, with and without genetic groups, and with and without genotypes. These four analyses were performed, first using all phenotypes, with the resulting EBVs for validation animals denoted as $\hat{\boldsymbol{u}}_w$. Phenotypes for the validation animals were then removed and the analyses repeated, with the resulting EBVs denoted as $\hat{\boldsymbol{u}}_p$. The subscripts "w" and "p" refer to "whole" and "partial" analyses respectively, with the partial EBVs of validation animals ($\hat{\boldsymbol{u}}_p$) informed through their pedigree and genomic relationships with the training animals. Maternal effects were fitted as routinely calculated in BREEDPLAN, but were not examined in the cross-validation, because the validation animals were not chosen to remove all phenotypes connected to the dam. EBVs were analysed on the underlying scale.

Correlations were used to examine the change in EBVs between each analysis. Cross-validation metrics were calculated using the method of Legarra *et al.* (2018). Traditional phenotype-based cross-validation metrics were not considered for this analysis due to the categorical nature of the calving ease trait. Accuracies were calculated by the formula

$$acc = \sqrt{\frac{cov(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p)}{(\overline{diag(\boldsymbol{K})} - \overline{\boldsymbol{K}})\sigma^2_{u,\infty}}}$$

where $\boldsymbol{K}$ is the appropriate relationship matrix for the validation animals with phenotypes for each trait and $\sigma^2_{u,\infty}$ is genetic variance in the validation animals, assumed to be the genetic variance. The dispersion was estimated by $disp = cov(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p)/var(\hat{\boldsymbol{u}}_p)$ and the bias was estimated as $bias = (\overline{\hat{\boldsymbol{u}}_p} - \overline{\hat{\boldsymbol{u}}_w})/\sqrt{\sigma^2_u}$, which was modified by Legarra *et al.* (2018) to allow for comparison between traits. While the validation animals included both genotyped and pedigree-only animals, metrics calculated only included genotyped animals due to computational difficulties. Metrics were also only calculated on direct effects, without consideration of maternal effects. Analyses were performed with the AGBU commercial solver on a computer with 2 x Intel(R) Xeon(R) E5-2697 v3 CPUs.

Table 1. Summary of the data used in the cross-validation studies

	Angus	**Hereford**
# animals in pedigree	3,006,655	2,247,767
# animals genotyped	200,259	34,585
# phenotypes		
Birth weight (BWT)	1,707,804	781,505
Calving difficulty score (CDS)	482,565	325,978
Gestation length (GL)	519,274	119,468
# validation animals with phenotypes		
Birth weight (BWT)	125,780	48,064
Calving difficulty score (CDS)	37,383	23,818
Gestation length (GL)	47,865	8,345
Proportion of CDS scores: 1,2,3	96.1, 2.7, 1.2	93.2, 4.7, 2.1

RESULTS AND DISCUSSION

A summary of the data used in the forward cross-validation is presented in Table 1. The correlation between EBVs from pedigree models with and without genetic groups for all animals

was 0.912 for both the Angus and Hereford analyses. When considering animals born from 2019 onwards, this correlation increased to 0.995 and 0.990 for the Angus and Hereford analyses, respectively. For the models without genetic groups, the correlations between pedigree and ssGBLUP models were 0.994 and 0.990 for the Angus and Hereford analyses, respectively. This decreased for the 2019-born animals to 0.886 and 0.961 for the Angus and Hereford, respectively. For recent animals most likely to be used for selection, inclusion of genomic information had a larger impact on changes in EBVs than inclusion of genetic groups.

Table 2. Cross-validation metrics for the Angus and Hereford analyses calculated based on genotyped animals born in 2019 or later

	EBV	n	Pedigree	Pedigree GG	Single-Step	Single-Step GG
			Angus			
Accuracy	BWT	45,613	0.475	0.475	0.840	0.839
	CE	14,606	0.340	0.340	0.533	0.534
	GL	19,351	0.441	0.442	0.672	0.676
Dispersion	BWT	45,613	0.983	0.982	1.030	1.029
	CE	14,606	0.997	0.999	1.025	1.026
	GL	19,351	0.941	0.950	0.992	0.995
Bias	BWT	45,613	0.002	-0.024	-0.002	-0.033
	CE	14,606	-0.013	-0.051	-0.010	-0.056
	GL	19,351	0.021	-0.029	0.001	-0.057
			Hereford			
Accuracy	BWT	10,285	0.677	0.672	0.869	0.863
	CE	5,715	0.401	0.413	0.516	0.526
	GL	2,670	0.555	0.646	0.655	0.718
Dispersion	BWT	10,285	0.968	0.965	1.010	1.012
	CE	5,715	0.942	0.917	1.008	0.992
	GL	2,670	1.149	1.019	1.127	1.052
Bias	BWT	10,285	-0.014	-0.003	-0.015	0.006
	CE	5,715	0.015	0.017	0.001	0.014
	GL	2,670	0.138	0.038	0.133	0.029

The forward cross-validation results for the Angus and Hereford analyses are presented in Table 2. For the Angus analyses, adding genetic groups to either pedigree or ssGBLUP models had virtually no impact on accuracy. Adding genomic information on the other hand improved accuracy substantially over pedigree-only analyses, by 0.365, 0.194, and 0.231 for BWT, CE and GL EBVs respectively in the ssGBLUP model without genetic groups. Little change was also observed in the dispersion, with all analyses close to the expected value of 1, indicating little evidence of over- or under-prediction. An increase in bias was observed for genetic group models for all traits, especially CE, with the bias increasing from -0.013 to -0.051 in the pedigree model, and from -0.01 to -0.056 in the ssGBLUP model.

For the Hereford analysis, the addition of genetic groups to the pedigree model increased accuracy for CE and GL EBVs, respectively, but as with the Angus analysis, adding genomic information had the largest impact on accuracy. Dispersion was improved for ssGBLUP models, with evidence for over-prediction in pedigree models (regressions < 1). The pattern of changes in bias was not consistent across traits and analyses, but for the CE trait itself, the ssGBLUP model without genetic groups had the least bias.

Based on these validation results, the inclusion of genomic information in ssGBLUP had a large benefit to prediction by increasing accuracies, and in some cases correcting dispersion and

minimising bias. Similar benefits were not apparent from the addition of genetic groups, which had no or minor benefit for accuracy and increased bias in CE EBVs for both analyses. Dispersion was largely unaffected by the model, but there was evidence for over-prediction for pedigree models in Herefords. While the pedigree accuracy for Angus is lower than Hereford, this is likely due to differences in the data structure for the two validation groups and warrants further investigation. While large increases in accuracies were observed for the genotyped validation animals, a smaller increase in accuracy is expected for the non-genotyped animals. Animals directly related to a genotyped animal will experience the greatest benefit from single-step, while animals less related will derive a lower benefit. It should also be noted that these validation metrics reflect the expected change for animals without a phenotype, and that individual animal results will vary. These results need to be verified for maternal effects but will require modifications to the validation set design.

Computation times for the models, including genomic information or genetic groups, had a large impact on the commercial viability of these analyses. For the Angus analyses, the model without genomics or genetic groups took 10,377 iterations to converge and 2.03 hours. The addition of genetic groups to this model required 19,986 iterations and 5.9 hours. The genomic model without genetic groups required 11,375 iterations and 20 hours to converge, while the addition of genetic groups increased this to 20,038 iterations and 37.34 hours. While the increase in computation times from the addition of genomic information is large, there is a corresponding increase in accuracies. The addition of genetic groups had no benefit to accuracies and almost doubled the number of iterations required. Therefore, inclusion of genetic groups constructed with the current strategy in this analysis is not recommended.

Calculating the mean of the $\boldsymbol{K}$ matrix for each trait makes using the Legarra *et al*. (2018) method challenging for pedigree-only animals when validating a single-step analysis. While an algorithm exists for calculating the diagonal of the ssGBLUP relationship matrix $\boldsymbol{H}$ (Legarra $et\ al.$ 2020), summary statistics for blocks of $\boldsymbol{H}$ are a challenge. For genotyped animals, the block of the $\boldsymbol{H}$ matrix required is a sub-matrix of the genomic relationship matrix $\boldsymbol{G}$, which can be calculated easily, but the other subblocks of $\boldsymbol{H}$ are more complex. One approach could be to solve the equation $\boldsymbol{v}'\boldsymbol{H}^{-1}\boldsymbol{v}$ by conjugate gradient, where $\boldsymbol{v}$ is a vector of zeros, except in the positions of the validation animals, which are set to $1/n$, where n is the number of validation animals.

CONCLUSION

Clear improvements in predictive ability were obtained for genotyped animals with the addition of genomic information in ssGBLUP models. However, the addition of genetic groups did not provide any improvements in calving ease direct predictions. Given the significant increase in computation time required to add genetic groups to the model, this term can be left out of the model without impact on recently born animals who are candidates for selection.

ACKNOWLEDGEMENTS

This research was partially funded by Meat and Livestock Australia project L.GEN.2204. The authors acknowledge the Agricultural Business Research Institute, Angus Australia and Herefords Australia for the provision of and access to the data for this study.

REFERENCES

Gilmour A.R. and Thompson R. (1998) *Proc. 6th World Congr. Genet. Appl. Livest. Prod*. **20**: 613.

Hoeschele I., Tier B. and Graser H.-U. (1995) *J. Anim. Sci.* **73:**1609.

Jeyaruban M.G., Johnston D.J., Tier B., Graser H.-U., Jeyaruban M.G., Johnston D.J., Tier B. and Graser H.-U. (2015) *Anim. Prod. Sci.* **56:** 927.

Legarra A., Aguilar I. and Colleau J.J. (2020) *J. Dairy Sci.* **103:** 3363.

Legarra A. and Reverter A. (2018) *Genet. Sel. Evol.* **50:** 53.