Validating the Accuracy of Genomic Evaluations for the Progeny of Unproven Sires

M. Sargolzaei[*†], J. P. Chesnais[‡] and F. S. Schenkel[*]

Introduction

Genomic evaluations have been officially implemented in Canada since August 2009 (Van Doormaal *et al.* (2009)). The genomic parent averages (GPA) of young males and females have lower accuracy than traditional proofs obtained from progeny testing. However, GPAs are available at the time of birth or soon thereafter while proofs are available when bulls are about five years old. Simulation studies have shown that substantial gains can be achieved by using genomic selection to turn the generations faster (e.g., Schaffer (2006)). However, these studies usually assume constant accuracy of GPA from one generation to the next. A simulation study by Habier *et al.* (2007) showed a significant drop in the accuracy of genomic breeding values (GEBV) estimated by a linear model when parents are unproven. The drop in accuracy was attributed to a lower relationship with the estimation group. The accuracy of GPAs of progeny of unproven sires is an important issue, because dairy cattle breeding companies are using a growing percentage of young unproven sires as sires of sons. In addition, if the drop in real accuracy of prediction over generations is large, then a larger number of proven bulls are required each generation to re-estimate SNP effects and restore the original accuracy. Therefore the issue has a direct impact on the total number of bulls that must be progeny tested either through organized progeny testing or through commercial sales of semen from young unproven bulls. Although theoretical reliabilities are routinely calculated for the GPA of sons of unproven sires, no study has been reported yet to validate the accuracy of prediction of these GPA based on real data. Therefore, the objective of this study was to examine the accuracy of the GPA of young bulls with an unproven sire via validation using real data.

Material and methods

Data. The data set consisted of the genotypes, pedigree information and traditional proofs of 31,044 Holstein bulls genotyped in North America using the Illumina BovineSNP50K BeadChip (Illumina Inc., San Diego, CA). SNP editing was performed as described in Wiggans *et al.* (2009) except for the minor allele frequency threshold which was set to 0.01 in this study. After editing, 43,385 SNPs were retained in the analysis. The pedigree data consisted of 103,963 animals with a cutoff point of 1950.

---

[*] Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, ON, Canada
[†] L'Alliance Boviteq, Saint-Hyacinthe, QC, Canada
[‡] Semex Alliance, Guelph, ON, Canada

For validation purposes, bulls were separated into an estimation set and a validation set. The estimation set included genotyped bulls for which progeny test proofs were available as of February 2005. It included approximately 4,400 bulls. For this estimation set, February 2005 Canadian proofs were used. But if a Canadian proof was not available, the MACE (multiple across-country evaluation) proof was used instead. Schenkel el al. (2009) showed that the use of MACE bulls can substantially increase the size of the estimation group and, as a consequence, increase the accuracy of GEBVs. The validation set included about 1,000 bulls that had a parent average, but no proof in February 2005, and had a Canadian proof in August 2009.

In order to mimic a situation where the sires of bulls in the validation group where unproven, the original estimation group was modified by excluding from it the sires of all bulls that were in the validation group. This ensured that the proofs of these sires were not used for SNP estimation. In total, 108 sires were removed. Both the original and modified data sets were analyzed. Once direct genomic values (DGV) were estimated from the modified estimation group, GPAs were calculated for all the other bulls, including the validation bulls and their sires. The GPA of validation bulls were then correlated with their daughter yield deviation (DYD) obtained from the August 2009 evaluation. The resulting correlations and reliabilities were compared to those obtained with the original estimation group, i.e. when the proofs of sires of validation bulls are used in the estimation of SNP effects.

Genomic evaluation models. Best linear unbiased prediction (BLUP) using a genomic relationship matrix was used to calculate DGVs in both the estimation and validation sets (VanRaden *et al.* (2009)). Since all genetic variation is not captured by the 50K SNP panel, a polygenic effect corresponding to 20% of the total genetic variance was included in the model. The DGVs were then blended with traditional proofs or PA to obtain GEBV or GPA, based on their respective reliabilities as described by Sullivan (2009). In the blending step, a 20% discount factor on DGV reliabilities was applied. Studies carried out at the Canadian Dairy Network (CDN) have shown that these values provide an optimal balance between prediction accuracy and obtaining an appropriate scaling of GPA compared to proofs.

Three alternate calculations of prediction accuracy were carried out for the validation bulls: 1) theoretical reliabilities (Rel_t), which were obtained from the inverse of mixed model equations 2) the squared correlation between 2005 GPA and 2009 DYD, and 3) realized reliabilities, which were calculated as: $Rel_r = (r^2_{GPA} / Rel_{DYD}) + (Rel_{PA} - (r^2_{PA} / Rel_{DYD}))$, where r^2_{GPA} and r^2_{PA} are the squared correlation between the 2005 GPA and the 2009 DYD, and between the 2005 PA and the 2009 DYD, respectively. Realized reliabilities take into account the accuracy of 2009 DYD and prior selection on parent average.

Seven key traits were considered: Milk yield, Fat yield, Protein yield (Prot), Fat%, Prot%, Somatic cell score (SCS) and Overall conformation (Conf).

Results and discussion

The squared correlations of the PA, DGV and GPA of validation bulls with their DYD are presented in Table 1, along with the realized and theoretical reliabilities of the GPA, when the sire was either proven or unproven. Correlation between the PA and DYD were the smallest as expected.

The squared correlations of DGV and GPA with DYD when the sires of validation animals were proven were in good agreement with those reported by Schenkel el al. (2009) and

Kistemaker and Sullivan (2009). While for proven sires scenario the squared correlations of GPA with DYD were identical or slightly smaller than those between DGV and DYD, they were all higher than those of unproven sires. As expected for all traits, the theoretical reliabilities of DGV and GPA were smaller for bulls with unproven sires than for bulls with proven sires. The same was true for realized reliabilities. The drop in realized reliabilities was equal or slightly smaller than the drop in theoretical reliabilities, which shows that theoretical reliabilities were slightly more conservative for the second generation than for the first, especially for SCS and conformation. Since the drop in the reliabilities of GPA in the second generation was relatively small, one could argue that the increased use of young sires as sires of sons would lead to more rapid genetic gain, since the reduction in generation interval more than compensates for the reduction in prediction accuracy. In addition, a relatively small number of progeny tested bulls may be required each year to balance the drop in accuracy.

Table 1: Squared correlations and reliabilities for validation bulls from proven (P) and unproven (U) sires

		r^2					Rel_t		Rel_r	
Trait	No.	PA	DGV_P	DGV_U	GPA_P	GPA_U	GPA_P	GPA_U	GPA_P	GPA_U
Milk	1014	0.23	0.48	0.39	0.47	0.43	0.61	0.55	0.65	0.59
Fat	1014	0.15	0.44	0.37	0.42	0.39	0.61	0.55	0.68	0.64
Prot	1014	0.20	0.44	0.38	0.44	0.40	0.61	0.55	0.64	0.60
Fat%	1014	0.21	0.57	0.51	0.53	0.51	0.61	0.55	0.73	0.70
Prot%	1014	0.20	0.49	0.41	0.48	0.45	0.61	0.55	0.69	0.66
SCS	1015	0.22	0.37	0.33	0.37	0.36	0.58	0.52	0.53	0.52
Conf	1016	0.20	0.40	0.32	0.40	0.36	0.57	0.51	0.59	0.55

PA: parent average; DGV: direct genomic value; GPA: genomic parent average;
Rel_t: theoretical reliabilities; Rel_r: realized reliabilities

The effect of having an unproven sire on regression coefficients of predictions over proofs was also investigated (Table 2).

Table 2: Ratio of regression coefficients of young bulls over proven bulls

Trait	DGV_P	DGV_U	GPA_P	GPA_U
Milk	0.77	0.74	0.92	0.94
Fat	1.00	0.95	1.16	1.18
Prot	0.77	0.76	0.91	0.93
Fat%	0.94	0.89	1.08	1.10
Prot%	0.86	0.80	1.05	1.09
SCS	0.76	0.72	0.91	0.94
Conf	0.80	0.77	0.94	1.00

DGV: direct genomic value; GPA: genomic parent average;

The regression of genomic predictions over the DYD of validation bulls was divided for each trait by the regression of bull proofs over subsequent DYD. The resulting ratio was compared when sires were either included or excluded from the estimation group (i.e. sire proven or unproven). When the amount of genotyping selection is similar for young and proven bulls, as was nearly true for this data set, the ratio should be equal to 1. If the ratio is less than one, the range of GPA will be over-estimated compared to proofs. If it is more than 1, their range will be under-estimated. From the Table 2, it appears that in general DGV tend to have too big a range compared to proofs, i.e. the GPA of top young bulls are over-estimated and those of bottom young bulls under-estimated. The same ratio for GPA is much closer to 1. When young bulls are from unproven sires, the ratio decreases, which shows that over-estimation in the range of the DGV gets worse. On the other hand, the ratio increases slightly for the GPA of bulls with unproven sires, leading to a more conservative range of GPA compared to proofs. In the Canadian system, the GPA therefore behaves significantly better than the DGV in terms of providing an appropriate range of predictions compared to proofs.

While this validation likely provides a fairly good assessment of the prediction accuracies when the sires are unproven, the results should be considered preliminary until a more complete study is carried out. In this study, the proofs of the sires of validation animals were removed by taking these sires out of the estimation group. Therefore, these proofs no longer affect the SNP effects directly. However, they could indirectly affect the proofs of contemporary bulls that are used in the estimation of SNP effects, since traditional proofs for both were estimated simultaneously. A better approach would then consist in cutting off all data at a specific time point two generations back, and examining the impact on the accuracy of prediction of GPA for validation bulls that have lost their sire's proofs. This would give a closer approximation of what occurs in practice when predictions are made two generations forward. With this new approach, it would also be worthwhile to compare relative drops in accuracy of GPA for linear and non-linear models, since according to Habier *et al.* (2007), the later should result in smaller decreases of accuracy over successive generations.

Conclusion

The use of young unproven bulls as sires of sons is growing. Preliminary results from this validation study show that the decrease in the prediction accuracy of genomic evaluations of young bulls is relatively small for bulls with unproven sires compared to bulls with proven sires and that the scale of the GPA is not adversely affected. Therefore, only a relatively small number of additional progeny tested bulls may be required each year to restore the original accuracy.

References

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). *Genetics*, 177: 2389–2397.

Kistemaker, G. J. and Sullivan, P. G. (2009). *Interbull Bulletin*. August 21–24.

Schaffer, L. R. (2006). *J. Anim. Breed. Genet.*, 123:218–223.

Schenkel, F. S., Sargolzaei, M., Kistemaker, G. *et al.* (2009). *Interbull Bulletin*.

Sullivan, P. G. (2009). *Technical Report to the CDN's Genetic Evaluation Board Meeting*, June.

Van Doormaal, B. J., Kistemaker, G. J., Sullivan, P. G. *et al.* (2009). *Interbull Bulletin*.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R. (2009). *J. Dairy Sci.*, 92:16–24.

Wiggans, G. R., Sonstegard, T. S., VanRaden, P. M. *et al.* (2009). *J. Dairy Sci.*, 92:3431–3436.