Improved genomic validation including extra regressions

P.M. VanRaden
U.S. Department of Agriculture, Agricultural Research Service, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705-2350, USA

Abstract

Genomic predictions (GEBVs) are often validated by predicting later deregressed conventional evaluations or daughter yield deviations (dEBVs or DYDs) from earlier GEBVs. Predicting later GEBVs from earlier GEBVs could be easier for the public to understand and to verify than standard validation and could be applied to single-step models where the GEBVs account for genomic preselection but the later dEBVs do not. Genomic validations could also predict deregressed GEBVs (dGEBVs) that include only the new information from the gain in reliability. Changes in genetic trend or rank can also be tested as in validation of conventional EBVs by including extra regressions such as on birth year, parent average (PA), or expected future inbreeding (EFI) from the earlier evaluation. The new validation can compute model squared correlations (R^2) that ideally should be high, indicating stable evaluations, and predict GEBV difference ($GEBV_{later} - GEBV_{earlier}$) to give residual R^2 that ideally should be low, indicating that changes in evaluations are not a function of other known factors or the earlier GEBV. The new validation methods were applied to U.S. GEBVs for 8 main traits. For most, the regressions on birth year indicated that genetic trend decreased as daughters were added, the regressions on PA were negative, indicating too much blending of PA with direct genomic value, the regressions on EFI were not significant, and regressions on earlier GEBV were >1.0 when the extra regressions were included. The model R^2 ranged from 48 to 79%, and the residual R^2 ranged from 3 to 18%. These new, more flexible methods give a more complete picture of GEBV properties and how models may be improved to reduce bias and improve prediction accuracy.

Key words: model validation, preselection bias, genetic trend, data truncation

Introduction

Countries participating in Interbull evaluations have validated their genetic and genomic models for many years. Trend tests 1, 2, and 3 (Boichard et al., 1995) can detect biases in estimates of genetic trend but are applied only to conventional estimated breeding values (EBVs) of proven bulls. Test 4 checks for stability of EBV or Mendelian sampling variance.

Genomic validation (Mäntysaari et al., 2010) can detect biases in the slope of response to selection but is only applied to test how well genomic EBVs (GEBVs) of young bulls match their later daughter yield deviations (DYDs). A new, combined validation could test both genetic trend and the slope of EBV or GEBV response for both young and old bulls in multi-step or single-step evaluations by including additional regressions, such as for birth year, when predicting final GEBV or deregressed GEBV (dGEBV) from earlier GEBV.

The standard Interbull genomic validation does not test birth year trend, proper blending of genomic with pedigree data, effects of inbreeding or other issues. Because later DYDs might be biased by genomic pre-selection, future options for genomic validation would be to use later GEBVs (Legarra and Reverter, 2018) or dGEBVs as the dependent variable instead of later DYDs or deregressed EBVs

(dEBVs). Such methods could give better results for testing single-step evaluations (e.g., ssGBLUP).

Materials and Methods

Prediction of later data should account for selection in earlier data. Conventional DYD excludes genomic information and could be biased. The later GEBV contains information from the earlier GEBV but is not independent. A simple dEBV is often created by separating parent average (PA) from progeny information to obtain a dependent variable

$$dEBV = PA + (EBV - PA)/REL_{diff},$$

where REL_{diff} is reliability (REL) calculated from the difference of total minus PA effective daughter contributions (EDCs). Similarly, a dGEBV can be created from the difference between earlier and later GEBVs:

$$dGEBV = GEBV_{earlier} + (GEBV_{later} - GEBV_{earlier})/REL_{diff}.$$

The RELs of $GEBV_{later}$ and $GEBV_{earlier}$ are both converted to EDCs, and then the difference of $EDC_{later} - EDC_{earlier}$ is converted back to REL to obtain REL_{diff}. The dGEBVs are weighted in the validation by this REL_{diff} computed from the difference in RELs. An animal's dGEBV gets no weight in the regression if the later evaluation had no gain in REL. Using dGEBV as the dependent variable helps to correctly estimate any biases. Those are often underestimated using $GEBV_{later}$ because of the part-whole relationship of $GEBV_{earlier}$ with $GEBV_{later}$ and REL < 1.0 (Macedo et al., 2020). Bulls with less accurate dGEBVs have more error variance and get less weight in the validation, whereas a correct weighting strategy is not clear if $GEBV_{later}$ was used as the dependent variable.

Gains in REL of dGEBV come from added parent information and higher REL_{PA}, genomic information from larger reference populations, own records for cows, daughter records as summarized in DYDs, and granddaughter records that might provide more information than daughters but are not included in DYDs. The validation tests the sum of all changes in GEBVs weighted by REL_{diff} obtained from all sources as a function of REL. Including other regressions such as on bull age can test if the genetic trend changes when the young bulls later add daughter records. Including regressions such as on PA or inbreeding can test if the blending of genomic and pedigree information is optimal or if bulls highly related to the population change more than others.

Regressions were computed using official GEBVs for 8 traits of 3,504 U.S. Holstein bulls with daughters in >10 herds in December 2020 but none in December 2016. The U.S. evaluations had a base change during that time as well as changes in the multi-trait estimation of productive life, fertility evaluation models, and the net merit formula. The base of the earlier GEBVs was adjusted to the later base using recent bulls that had little change in REL during the 4 years. The base adjustment for each trait used bulls born since 2005 that had ≥500 daughters and an REL of 97% in 2016. Numbers of bulls to change the base ranged from 845 for yield traits to only 30 for heifer conception rate. More flexible edits are needed for smaller breeds or populations.

U.S. evaluations adjust for inbreeding because it affects many traits (VanRaden, 2005). Thus, GEBVs may change for bulls with an average relationship to the breed (EFI) that increases during the 4 years between earlier and later data because they, their sire, or their grandsires may contribute much DNA to the breed. While predicting 2020 GEBV from 2016 GEBV, a regression was also included on the bull's 2016 EFI to measure changes in GEBV associated with inbreeding. In other countries that have not adjusted for inbreeding, nonadditive genetic effects of inbreeding and changing populations of mates could also explain changes in EBVs or GEBVs across time and regressions that differ from 1.0.

To allow simpler comparisons of intercepts (b_0) and regressions (b_1) across traits and factors, t-test values are presented instead of standard errors or probabilities of a larger t-value. The unitless t-values preserve the direction and magnitude of the regressions but are approximate because changes for family members may be correlated. Values less than ±2 are statistically insignificant; values larger than ±2 are significant ($P < 0.05$), but much higher values might be required for biological significance because many bulls were included. Also, the t-test values for b_1 check if the regression on previous GEBV is statistically different from 1.0 rather than different from 0.

Squared correlations (R^2) are presented in 2 ways. Use of dGEBV as the dependent variable gives model R^2, whereas use of the dGEBV difference ($dGEBV_{later} - GEBV_{earlier}$) gives residual R^2. Ideally the model R^2 should be high, indicating stable evaluations, but the residual R^2 should be low, indicating that changes in evaluations are not a function of other known factors or the earlier GEBV. Thus, GEBV changes should not be predictable, and earlier GEBVs should be adjusted to the later genetic base to make comparisons meaningful.

Results & Discussion

Simple regressions that used different dependent variables gave similar b_0 and b_1 but much different R^2 because of the information included (Table 1). The regression b_1 was near 1.0 without extra regressions as in standard genomic validation. The DYD gave lowest model R^2 by predicting only new daughter records, whereas dGEBV had higher R^2 by including information from all new phenotypes and genotypes; GEBV had highest R^2 by also predicting earlier information.

Table 1. – Comparison of dependent variables used for validating genomic predictions of milk yield

Dependent		t-test		R^2 (%)	
variable	b_1	b_1[a]	b_0	Residual	Model
DYD	1.03	1.9	−14.4	0.10	59
dGEBV	0.99	−0.9	−16.5	0.02	69
GEBV	0.99	−0.8	−16.3	0.02	72

[a]Test of b_1 difference from expected 1.0.

The 3 extra regressions were each significant whether fit separately or together (Table 2). Later dGEBV declined for the youngest bulls, those with highest PA, and those with higher EFI. The b_1 became much greater than 1.0 when extra regressions were added. As an example, the b_1 for milk yield with only birth-year regression added increased to 1.08 (0.08 higher than the expected 1.0) and had a standard error of only 0.03 (not shown), giving a highly significant t-value of 7.1 for b_1 difference from 1.0. Without the extra regressions, the regression of later dGEBV on earlier GEBV for milk yield was 0.99 with a model R^2 of 69%. With all 3 terms added, the model R^2 increased to 77% from 69% in the simple validation.

Regressions for 7 other traits from the validation model with 3 extra regressions are presented in Table 3. The b_1 are >1.0 for most traits, and corresponding PA regressions are negative because U.S. GEBVs have put extra weight on traditional PA or EBV to reduce overestimation in the youngest birth years. The b_0 are near 0 for yield traits but larger for some other traits because for example the fertility models were revised during those 4 years. The birth year regressions were negative for yield traits, indicating that genetic trend estimates declined with additional later data, but were small for most other traits. Those regressions for yield may imply that trend is too high in youngest bulls or is too low in progeny-tested bulls because of preselection bias. Similar regressions and R^2 gains were obtained in Canada when birth year and PA regressions were included for yield.

Regressions on EFI were all negative, indicating that GEBVs of the most popular bulls decreased more than expected. In Table 3, model R^2 was higher for traits with higher heritability such as yield traits, and residual R^2 was smaller for less heritable traits. The residual

Table 2. – Comparison of extra regressions for bull birth year (as difference from earliest year), PA, and EFI included separately or together for milk yield

	t-test (significance)					R^2 (%)	
b_1	b_1 [a]	Birth year	PA	EFI	b_0	Residual	Model
0.99	−0.9	…	…	…	−14.4	0.02	69
1.08	7.1	−19.1	…	…	−25.7	9	72
1.43	19.4	…	−22.6	…	−14.7	13	73
1.00	−0.2	…	…	−4.6	3.4	1	69
1.47	20.0	−14.8	−18.5	−3.1	1.0	18	77

[a]Test of b_1 difference from expected 1.0.

Table 3. – Genomic validations predicting dGEBVs for several traits including extra regressions on bull birth year (as difference from earliest year), PA, and EFI

		t tests (significance)					R^2 (%)	
Trait	b_1	b_1 [a]	Birth year	PA	EFI	b_0	Residual	Model
Milk yield	1.47	20.0	−14.8	−18.5	−3.1	1.0	18	77
Fat yield	1.40	18.7	−12.9	−18.6	−0.8	−0.7	16	75
Protein yield	1.47	17.5	−14.4	−16.3	−3.1	1.8	13	73
Somatic cell score	1.22	11.6	1.4	−3.7	−4.1	−4.6	6	73
Productive life	0.89	−3.8	−6.9	4.1	−6.1	4.0	3	53
Daughter pregnancy rate	0.91	−3.9	−14.1	3.8	−0.8	−2.4	6	58
Cow conception rate	1.15	7.0	−11.3	−3.9	−1.5	−2.0	4	60
Heifer conception rate	1.10	3.3	−6.0	6.0	−6.5	4.4	5	48

[a]Test of b_1 difference from expected 1.0.

R^2 sums the variance for the 3 regressions: birth year, EFI, and b_1 difference from 1.0. The prediction correlations were very good for all traits but lower for heifer conception rate because of a smaller reference population and lower heritability.

Previous efforts to pass Interbull validation may have restricted the GEBVs of young bulls too far for some traits. The previous blending of the direct genomic value with PA had reduced b_0 bias but also reduced R^2. As a result, rankings may be less accurate than possible for young bull predictions, and breeders may have shifted back to using progeny-tested bulls more than deserved if GEBVs of top young bulls are underestimated.

Based on these results, the weight on traditional EBVs for yield traits was reduced in the U.S. selection index blending from the previous 15% down to 10% for several traits of Holsteins in August 2021. This adjustment does not affect the marker effects but is an option in a later program that does the blending. Reducing this weight has very little effect on progeny-tested bulls because of small differences between their traditional and genomic EBVs but gives more weight to the direct genomic values of young animals.

Conclusions

Validation could use published GEBV or dGEBV and ssGBLUP. Predictions of GEBV are simple to explain but not independent. Later dGEBVs are independent of earlier GEBVs. Extra regressions can show which bull groups change and why. Trend differences may reflect inflation of GEBVs for the youngest bulls or preselection bias in GEBVs for progeny-tested bulls. Models may need revision to balance accuracy and bias.

Acknowledgments

The Council on Dairy Cattle Breeding (Bowie, MD) provided the data for this research under USDA Agricultural Research Service (ARS) Material Transfer Research Agreement

58-8042-8-007. The author thanks Esa Mäntysaari, Pete Sullivan, Raphael Mrode, and Zengting Liu for their helpful suggestions as members of Interbull's Validation Working Group with coordination by Valentina Palucci. Funding for P.M. VanRaden was from USDA-ARS appropriated project 8042-31000-002-00-D, "Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals." USDA is an equal opportunity provider and employer.

References

Boichard, D., Bonaiti, B., Barbat, A., Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci. 78,* 431-7.

Legarra, A., Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol. 50,* 53.

Macedo, F.L., Reverter, A., Legarra, A. 2020. Behavior of the Linear Regression method to estimate bias and accuracies with correct and incorrect genetic evaluation models. *J. Dairy Sci. 103,* 529-544.

Mäntysaari, E., Liu, Z., VanRaden, P. 2010. Interbull validation test for genomic evaluations. *Interbull Bull. 41,* 17–22.

VanRaden, P.M. 2005. Inbreeding adjustments and effect on genetic trend estimates. *Interbull Bull. 33,* 81–4.