Accuracy and Bias of Genomic Prediction for Second-Generation Candidates

Z. Liu, H. Alkhoder, F. Reinhardt, and R. Reents

*IT-Solutions for Animal Production (**vit**), Heinrich-Schröder-Weg 1, 27283 Verden, Germany*

Abstract

As a result of intensive genomic selection in Holstein breeds, the distance of selection candidate to genomic reference population has increased. Most genotyped candidates nowadays have no sire with daughters in milk yet at the time of being selected for breeding, they are referred to as second-generation candidates. Genomic model up to now has been optimized for genomic prediction of candidates with their sire or dam in reference population, referred to as first-generation candidates, and has not accounted for the breakdown of linkage disequilibrium from first- to second-generation candidates. To quantify the loss in accuracy and the bias of genomic prediction for the second-generation candidates, a special genomic validation was conducted, based on genotype and phenotype data from the December 2015 genomic evaluation for German Holstein. As a comparison, a regular genomic validation was done based on the same validation bulls by treating them as first-generation candidates. Accuracy of genomic prediction of direct genomic values, shown in observed R^2 values of Interbull GEBV Test, was significantly lower for second-generation than first-generation candidates, and the decrease in R^2 values from first- to second-generation ranged from 0.02 to 0.14 with a mean of 0.086 for 37 MACE traits. A similar drop in the R^2 value was found also in conventional pedigree index. Bias of genomic prediction, expressed as ratio of regression slopes between the two validation scenarios, deviated also from its expectation. Variance of direct genomic values of the second-generation candidates was too high, in relation to that of the first-generation candidates, with an average of the ratio being 0.95 across all the 37 traits. A shrinkage factor for SNP effect estimates was proposed for direct genomic values in order to reduce the over-prediction for the second-generation candidates. By doing so, the same set of SNP effect estimates can be used for differentiated prediction of genomic breeding values for both the first- and second-generation candidates. The genomic model for German Holstein has been optimized for properly predicting genomic breeding values of second-generation candidates and the optimized model was introduced in April 2016.

Key words: genomic prediction, second generation candidates, accuracy, bias

Introduction

Since the implementation of genomic selection and evaluation for Holsteins in 2008 (VanRaden 2008), generation intervals have been reduced significantly, particularly for the pathway from sire to son shortened from about 6 to 2 years. The increased use of young genomic bulls as sires of the next generation animals has resulted in a widening gap between genotyped candidates and genomic reference population (RP). Genotyped candidates without sire in the genomic RP, which are referred to as second-generation candidates in this paper, have increasingly dominated the selection of breeding animals. Even grandsires of some animals or genotyped embryos have no daughters or are not included in genomic RP. Those candidates with neither grandsire nor sire in reference population are referred to as third-generation candidates thereafter. As a result of the intensive genomic selection, first-generation candidates with sire included in RP are disappearing. In routine genomic evaluation for German Holsteins, the second- or third-generation candidates dominated the top ranking list in the last years.

The genomic model (Meuwissen *et al.*, 2001) relies on the linkage disequilibrium (LD) between SNP markers and genes/mutations responsible for the inheritance of evaluated traits. When some of the LD are broken down from one to next generation, accuracy of genomic prediction will decrease and bias of estimated genomic breeding values (GEBV) will increase for new candidates. By fitting a residual polygenic effect (RPG, Liu *et*

al., 2011), the LD breakdown between RP and first-generation candidates can be accounted for. Up to now, the LD decay from the first to second/third generation has not been accounted for properly in genomic prediction. In addition, the current Interbull genomic validation test (Mäntysaari *et al.*, 2010) does not consider the LD breakdown from first to later generations either. The objectives of this study were to quantify the loss in accuracy and the bias of genomic prediction for the second-generation candidates and to develop a method for optimal genomic prediction for the second-generation candidates.

Materials and Methods

Distance between reference population and selection candidates

Tables 1, 2 and 3 show percentages of Holstein candidates without sire in EuroGenomics RP for German Holstein genomic evaluations in August 2011, August 2014 and December 2016, respectively. Three groups of traits are represented in the tables, milk yield as a regular trait, longevity as a late-measured trait, and direct effect of calving ease as an early-measured trait. Because Germany did not submit calving trait EBV to MACE evaluation in 2011, there was no corresponding information on direct calving ease in Table 1. Last three birth years from the year of genomic evaluation were chosen for the analysis. It can be clearly seen that the percentage of candidates with no sire in RP has increased from 23% for the regular trait milk yield in August 2011 to 92% in December 2016 genomic evaluation. An even higher percentage of candidates do not have sire in RP for the late-measured trait longevity than for the regular trait milk yield. In contrast, the early-measured trait, direct calving ease, has fewer candidates without sire in reference population. About 10-15% of the candidates without sire in RP have neither maternal nor paternal grandsire in the RP either; they are the so-called third-generation candidates. Comparing Tables 2 or 3 to Table 1, we can see that the first-generation candidates are disappearing and in contrast the later generation candidates are dominating the selection of breeding animals.

Table 1. Percentages of candidates without sire in EuroGenomics RP for German Holstein genomic evaluation in August 2011.

Year of birth	Number of candidates	Regular traits	Late traits	Early traits
2009	6,069	3%	10%	
2010	10,109	11%	20%	
2011	5,194	23%	39%	

Table 2. Percentages of candidates without sire in EuroGenomics RP for German Holstein genomic evaluation in August 2014.

Year of birth	Number of candidates	Regular traits	Late traits	Early traits
2012	32,441	34%	55%	0%
2013	33,870	78%	84%	3%
2014	15,361	91%	93%	24%

Table 3. Percentages of candidates without sire in EuroGenomics RP for German Holstein genomic evaluation in December 2016.

Year of birth	Number of candidates	Regular traits	Late traits	Early traits
2014	42,859	40%	80%	0%
2015	42,214	85%	92%	34%
2016	26,935	92%	94%	71%

A regular genomic validation for the first-generation candidates

Genotype and phenotype data for conducting a genomic validation originated from a routine genomic evaluation for German Holstein in December 2015. A total of 33,436 Holstein bulls were present in the EuroGenomics reference population for milk yield. According to the GEBV Test procedure (Mäntysaari *et al.*, 2010), 29,917 bulls born in and before 2008 were chosen as reference animals; whereas 1,063 younger German national bulls were treated as validation animals. Figure 1 shows the set-up of reference and validation bulls for the regular genomic validation by treating the validation bulls as first-generation candidates.

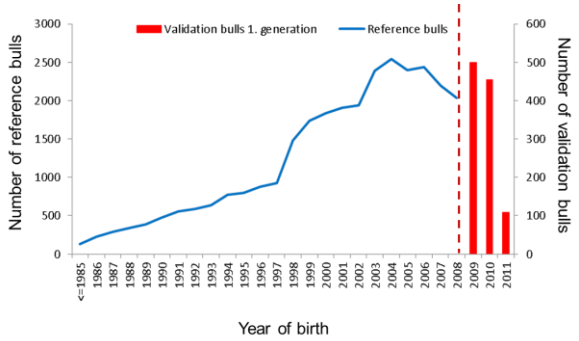

Figure 1. Reference and validation bulls for a regular genomic validation for first-generation candidates (Trait: milk yield).

In order to simulate second-generation candidates, sires, maternal grandsires (MGS) and paternal grandsires (PGS) of the validation bulls were traced back and their distributions across birth years were shown in Figure 2.

The 1063 validation bulls were linked to 178 sires, 137 MGS or 81 PGS. 734 of the 1063 validation bulls had sire born after 2002, whereas the rest 329 validation bulls had sire born in and before 2002. In order to use exactly the same validation bulls as first-generation and second-generation candidates in two validation scenarios, the 329 validation bulls with older sire were discarded, leaving 734 validation bulls with sire born after 2002 (Figure 3).

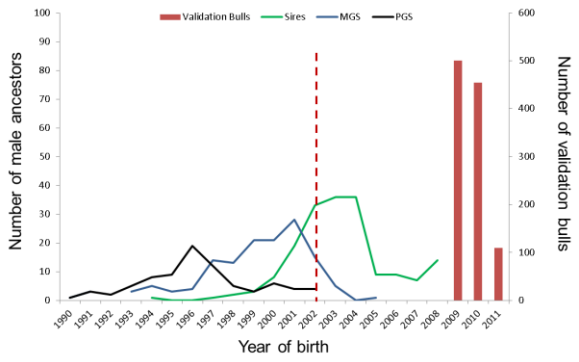

Figure 2. Number of sires, maternal and paternal grandsires of the validation bulls born in different years (Trait: milk yield).

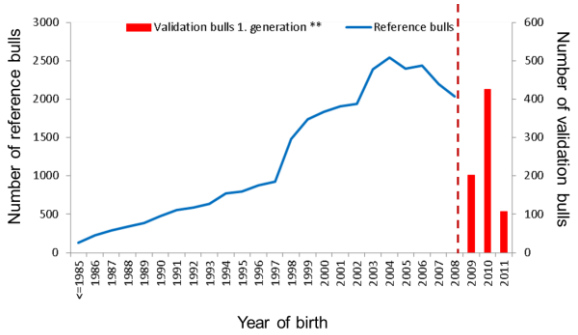

Figure 3. Final validation bulls selected as first-generation candidates with sire born after 2002.

A genomic validation for the second-generation candidates

The cut-off year of birth for reference bulls was set to 2002 for the genomic validation by treating the 734 validation bulls as second-generation candidates (Figure 4). Sires of the validation bulls were too young to be included in the RP containing 15,912 reference bulls. However, MGS or PGS of the validation bulls were present in this RP. Due to the history of genotyping of the Holstein reference bulls, the number of reference bulls in the RP for second-generation candidates, 15,912, was much smaller than the RP for first-generation candidates, 29,917. Nevertheless, the 15,912 reference bulls mostly with many daughters should result in reasonably accurate genomic prediction.

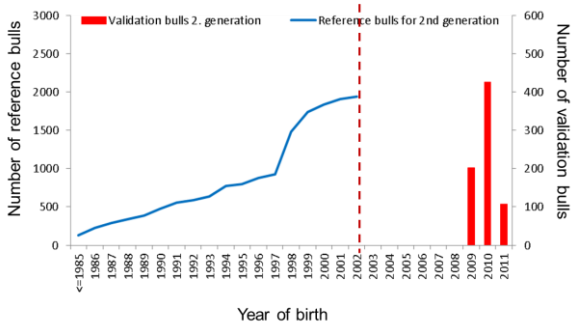

Figure 4. Reference and validation bulls for a genomic validation for second-generation candidates (Trait: milk yield).

Results & Discussion

SNP marker effects were estimated using the reference populations for each scenario of the genomic validation, by treating the same validation bulls as first- or second-generation candidates. All 37 traits included in MACE evaluation for German Holsteins were analyzed, except mastitis. Complete genomic evaluations were conducted for the two scenarios. Interbull's python software GEBVtest was used to perform the genomic validation test. Instead of GEBV as independent variables, direct genomic values (DGV) were used for the genomic validation test for first-/second-generation candidates:

$$DRP = b_0 + b_{1\,first} DGV_{first} + e \qquad [1]$$

$$DRP = b_0 + b_{1\,second} DGV_{second} + e \qquad [2]$$

where DRP represents deregressed EBV of the validation bulls, and the subscript *first* or *second* denotes the validation bulls being treated as first- or second-generation candidates. The observed R^2 values for the two regression models [1] and [2] are denoted as R^2_{first} and R^2_{second}, respectively. The EBV sub-model of the Interbull GEBVtest is for the both scenarios:

$$DRP = b_0 + b^{EBV}_{1first} EBV_{first} + e \qquad [3]$$

$$DRP = b_0 + b^{EBV}_{1\,second} EBV_{second} + e \qquad [4]$$

where EBV represents male pedigree index of the validation bulls. The observed R^2 values of the regression models [3] and [4] are denoted as $R^2_{first\_EBV}$ and $R^2_{second\_EBV}$, respectively.

Accuracy of genomic prediction

Figure 5 summaries the reduction in validation accuracy (observed R^2 value) of DGV by treating the validation bulls as second-generation compared to first-generation candidates. In general, second-generation candidates had lower accuracy than first-generation candidates. The reduction in the validation R^2 values ranged from 0.02 to 0.14, with an average of 0.086.

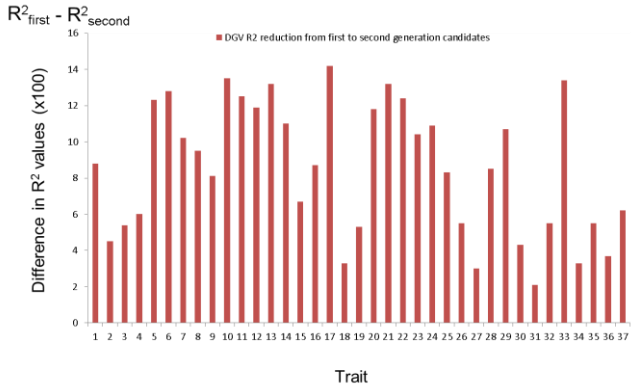

Figure 5. Reduction in validation R^2 values of DGV by treating validation bulls as second- to first-generation candidates.

As for DGV, Figure 6 shows the reduction in validation R^2 values of the regression models [3] and [4] for the male pedigree index. The decrease in R^2 values from first- to second-generation candidates ranged from 0.01 to 0.27, with an average of 0.065. The trait 23 with the largest R^2 decrease had a trait definition change in the past years.

Because the second-generation candidates are one generation more apart from bulls with phenotypes than the first-generation candidates, model reliabilities of the second-generation candidates are expected to be half of those of first-generation candidates. This can be clearly seen in Figure 7. Model reliabilities of male pedigree index were reduced from 0.34 to 0.17 averaged over all the traits. The changes in regression slopes of models [3] to [4] are less consistent across the traits than for DGV and have a larger variation among the traits.

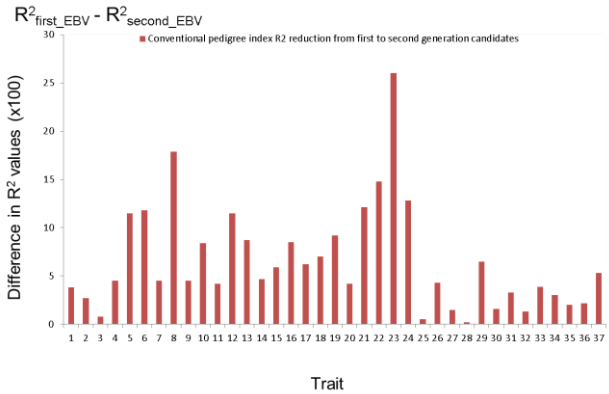

Figure 6. Reduction in validation R^2 values of male pedigree index by treating validation bulls as second- to first-generation candidates.

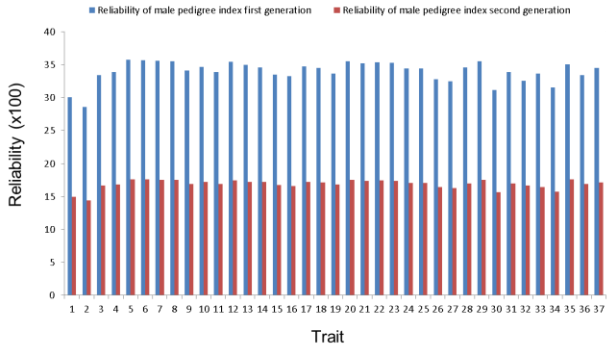

Figure 7. Model reliability values of male pedigree index for the validation bulls as first- or second-generation candidates.

The loss in accuracy of genomic prediction, shown in the observed R^2 values, was substantial for the second-generation candidates, in comparison to the first-generation candidates. The 734 validation bulls born in 2009 to 2011 were genomically pre-selected with varying selection intensity, thus the validation accuracy of the two scenarios might be influenced by the pre-selection. However, the difference in the observed R^2 values between the first- and second-generation candidate scenarios should be free of the impact of pre-selection of the validation bulls, because the same validation bulls were used in the two validation scenarios. The reduction in validation accuracy varied among all the 37 traits, indicating different selection pressures on the traits.

In general, smaller reference populations tend to result in lower validation R^2 values than larger reference populations. It can be argued whether the lower validation accuracy of the second-generation validation scenario be mainly caused by the smaller size of reference population rather than by a larger distance between the validation to the reference population. However, according to a simulation study by Interbull Genomic Reliability Working Group (unpublished data), second-generation candidates also had significantly lower accuracy than first-generation candidates even with identical reference populations for the two generations of candidates. The magnitude of the accuracy loss was similar as found in this study.

Bias of genomic prediction

Regression slope, b_1, of the genomic validation models [1] and [2] measures if the variance of DGV was too high, if $b_1 < E(b_1)=1$, or too low, if $b_1 > E(b_1)=1$. The ratio of the regression slopes of the second- to first-generation candidate scenarios:

$$f = b_{1\text{second}} / b_{1\,first} \qquad [5]$$

indicates the DGV standard deviation of the validation bulls being second-generation relative to the first-generation candidates. Figure 8 displays the ratios of the regression slopes for all the 37 traits.

Average of the ratio f is 0.95 for all the 37 traits, suggesting that DGV of the validation bulls as second-generation candidates has too high variance than being first-generation candidates. The relative lower regression slope for second-generation candidates, in comparison to first-generation candidates, was also found in the simulation study by the Interbull Genomic Reliability Working Group (unpublished data). The traits with the ratio being close to 1 or even higher, e.g. traits 14-16, 26 and 27, belong to either recently introduced traits or those trait definitions changed in the past years. Based on the ratio of regression slopes, we can draw a conclusion that variance of DGV of the second-generation candidates was too high in relation to the first-generation candidates.

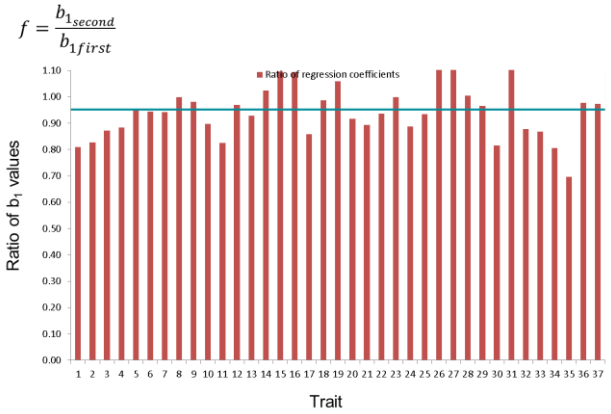

Figure 8. Ratio of regression slope estimates of the second-generation to first-generation candidates.

Implementation in routine evaluation

A majority of breeding animals of the Holstein breed are selected nowadays based on genomic evaluation at a time when their sires still have no daughters' phenotype records. Accurate genomic prediction of the second-generation candidates plays an important role for breeding. Therefore, the current genomic model for German Holsteins (Liu *et al.*, 2010) has been modified to properly account for candidates of different generations apart from genomic reference population.

Based on the EuroGenomics bull reference population, generation number is determined for each genotyped animal, on a trait-by-trait basis, in German Holstein genomic evaluation. Due to the much lower proportion of third-generation candidates, they are treated as if they belong to the second generation. Male pedigree index is calculated using deregressed MACE EBV of all Holstein bulls included in MACE evaluation as before, with reduced variance and reliabilities for second-generation than for first-generation candidates. DGV of a genotyped animal is computed as:

$$DGV = \hat{\mu} + f \sum_i z_i \hat{a}_i \qquad [6]$$

where $\hat{\mu}$ is estimated mean effect of reference population in SNP effect estimation, z_i is genotypic value (VanRaden, 2008) for SNP *i*, $\hat{a}_i$ is estimate of the *i*-th SNP marker, and *f* is a shrinkage factor for DGV:

$$f = b_{1\mathrm{second}} \Big/ b_{1\,first}$$

for second-generation candidate; otherwise

$$f = 1.$$

The average shrinkage factor, $f = 0.95$, for second-generation candidates across all the traits corresponds to 2.5% less DGV variance than the first-generation candidates. When DGV of the validation bulls in the second-generation validation scenario was calculated using formula [6], all the traits passed the Interbull genomic validation test via the GEBV Test software. The shrinkage factor on DGV enables the use of the same set of SNP effect estimates for calculating DGV of the first- as well as second-generation candidates.

This model optimization for the second-generation candidates was introduced in routine genomic evaluation for German Holstein in April 2016.

Conclusions

Intensive genomic selection has led to almost disappearing of first-generation candidates, mainly due to the shortened generation interval in the sire to son pathway. When cows are also included in genomic reference population, the dominance of second- or third-generation candidates still persist, if a minimum amount of phenotype data is required for cows entering the reference population, e.g. with at least one complete lactation. Genomic model relies on the LD between SNP markers and genes or mutations responsible for evaluated traits. The breakdown of LD from first- to second-generation or second- to third-generation was not accounted for in the current genomic evaluation. To assess the impact of the LD decay, a special genomic validation was conducted by treating validation bulls as second-generation candidates. As a comparison, the same validation bulls were used as first-generation candidate in the other genomic validation, which is the current standard procedure for testing national genomic evaluation. Comparing to the first-generation candidate scenario, accuracy of DGV for the second-generation candidates was reduced with an average decrease in observed R^2, 0.086, for all the analyzed 37 traits. Bias in DGV of the second-generation candidates, expressed as the ratio of regression slope to the first-generation candidates, was shown to be increased. The average ratio of the regression slope estimates was 0.95 for DGV of all the evaluated traits, suggesting an overestimation of DGV standard deviation of second-generation candidates, 2.5%, in comparison to the first-generation candidates. Fitting a residual polygenic effect in the estimation of SNP effects can account for the incomplete LD between SNP markers and genes and the LD decay from reference population to the first-generation candidates. A shrinkage factor for

DGV can consider the LD decay between first- and second-generation candidates, reducing the over-prediction of GEBV for second-generation candidates. The current Interbull GEBV Test, designed for first-generation candidates, might be extended to second-generation candidates too, by using the shrinkage factor for DGV of second-generation candidates.

Acknowledgements

Colleagues of the Interbull Genomic Reliability Working Group are kindly thanked for a simulation study on genomic reliability calculation for candidates of different generations apart from reference population.

References

Liu, Z., Seefried, F.R., Reinhardt, F., Rensing, S., Thaller, G. & Reents, R. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet. Sel. Evol. 43:19.*

Mäntysaari, E., Liu, Z. & VanRaden, P.M. 2010. Interbull Validation Test for Genomic Evaluations. *Interbull Bulletin 41,* 17-21.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps *Genetics 157:4,* 1819–1829.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions *J. Dairy Sci. 91:11,* 4414-4423.