A Simple Method for Correcting the Bias Caused by Genomic Pre-Selection in Conventional Genetic Evaluation

Z. Liu, F. Seefried, F. Reinhardt and R. Reents
vit *w.V., Heideweg 1, 27283 Verden, Germany*

Abstract

With the development of genomic evaluation in dairy cattle, an ever increasing number of countries have been pre-selecting genotyped calves as parents of animals of future generations with a much higher selection intensity than in the past. If this genomic pre-selection is ignored in conventional genetic evaluation, proofs will be biased. We developed a simple method to correct the pre-selection bias in conventional genetic evaluation. In order to validate our bias correction method, we simulated genomic pre-selection four years ago with three levels of culling rate and conducted test genetic evaluations with and without adjusting for the pre-selection bias. A total of 655 genotyped Holstein bulls, born in 2004, were treated as if they were genotyped selection candidates four years ago. Their official milk yield proofs from August 2009 German official conventional evaluation were removed from further analyses and these bulls were selected or culled based on their estimated direct genomic values. Results of the test evaluations were compared to the reference bull evaluation with about 24,000 Holstein bulls included. For all levels of culling rate, the simple bias correction method gave much smaller averages and standard deviations of proof biases. In particular, proofs of dams of the culled genotyped bulls were extremely overestimated when the pre-selection was ignored in the test evaluations. Overall, the proposed bias correction method seemed to be very efficient. Some aspects on further development were also discussed. The genomic pre-selection problem exists for international bull comparison, too. The proposed method can be easily implemented in both national and international genetic evaluations.

1. Introduction

More and more countries have implemented genomic evaluation and selection in dairy cattle. Genomic selection (Meuwissen *et al.*, 2001) based on massive marker information, e.g. single nucleotide polymorphism, can increase accuracy of pre-selection of breeding animals significantly. However, the pre-selected genotyped animals usually have higher genetic levels than randomly selected candidates as parents of next generations. If the genomic pre-selection is not accounted for in conventional evaluation, genetic evaluation is very likely biased. Patry and Ducrocq (2009) justified the existence of such a bias using a simulation study. The objectives of this study were to develop a method for adjusting the genomic pre-selection bias, and to validate the method with real data from an official evaluation.

2. Materials and Methods

2.1. The bias correction method

The proposed method for correcting genomic pre-selection makes two assumptions about estimated direct genomic values (DGV) and their associated genomic effective daughter contribution (EDC):

a) DGV of all genotyped animals, including culled ones, are available, and

b) Estimates of DGV and genomic EDC are unbiased.

Note that the genomic EDC refers to extra EDC contributed by genomic information and measures the information gain by using genomic relationship rather than average relationship among genotyped animals (Reinhardt *et al.*, 2009). In fact, these two assumptions are usually made in routine genomic evaluations.

For each genotyped animal, a pseudo-record is generated with:

$$q = \hat{\mu} + (\hat{a} - \hat{\mu}) / R^2_{\hat{a}} \qquad [1]$$

where q represents the generated record which is analogue to deregressed proof, $\hat{\mu}$ is estimated general mean of genotyped animals in reference population (Liu *et al.*, 2009), $\hat{a}$ is estimated DGV, sum of all SNP effect estimates, for the animal, and $R^2_{\hat{a}}$ denotes reliability of $\hat{a}$. A weighting factor for q is equal to its associated genomic EDC, which is calculated as in Ducrocq and Liu (2009):

$$\psi = \alpha\left(\frac{R^2_{\hat{a}}}{1 - R^2_{\hat{a}}} - \frac{R^{2\,[E3]}_{\hat{a}}}{1 - R^{2\,[E3]}_{\hat{a}}}\right) \qquad [2]$$

where ψ is the pure genomic EDC as weighting factor, α is the ratio of residual to sire variance, and $R^{2\,[E3]}_{\hat{a}}$ represents a reliability obtained from a subset conventional evaluation (VanRaden, 2008), denoted as E3, including only genotyped animals.

For genotyped animals with phenotypes, the pseudo-record is a weighted average of deregressed proof from conventional evaluation and q from Formula 1:

$$q_{comb} = (\psi\, q + \psi_C\, q_C) / (\psi + \psi_C) \qquad [3]$$

where q_{comb} denotes combined deregressed proof, q_C is deregressed proof from conventional evaluation, and ψ_C is EDC from conventional evaluation. Weighting factor for q_{comb} is sum of both EDC: $\psi + \psi_C$.

All genotyped animals, including culled ones, received a pseudo-record and corresponding genomic EDC. These records were added to regular conventional BLUP evaluation for all animals. As claimed by Ducrocq and Liu (2009), this BLUP approach represents a more accurate way of combining genomic and conventional information than the standard selection index on an animal by animal basis.

2.2. Data materials

2.2.1. Choice of a reference evaluation

Deregressed proofs of milk yield were computed for all 23,557 Holstein bulls with at least 10 EDC from August 2009 German national evaluation. In the deregression step sire-dam pedigree was used, instead of sire-maternal grandsire pedigree. A single trait model, including only a general mean and bull additive genetic effect, was fitted to the deregressed proofs to estimate breeding values. This bull evaluation considered sire-dam pedigree and ancestors were traced back as far as possible in pedigree, resulting in a total of 76,762 animals in final pedigree. This bull evaluation was compared to the official cow evaluation that included more than 22 million animals with a random regression test-day model (Liu *et al.*, 2004). Because no systematic biases were observed for all the bulls with daughters and their ancestors, and also because full scale cow evaluations would require much more resources, we decided that this bull evaluation be treated as an accurate substitute of national cow evaluation for this study and this particular bull evaluation with all bulls included as a reference evaluation for all test evaluations.

2.2.2. Simulating genomic pre-selection of different selection intensity

In the bull reference evaluation 655 bulls born in 2004 were genotyped. Their DGV estimated with a genomic model assuming equal SNP marker variance (Liu *et al.*, 2009; Reinhardt *et al.*, 2009) and associated reliabilities were used as the basis for genomic pre-selection that would have been made four years ago. Pseudo-records were calculated using Formulae 1 and 2. A total of 4339 genotyped training bulls were considered in the particular genomic evaluation, including all the youngest bulls born in 2004. Three levels of culling rate were simulated for pre-selecting the genotyped 2004 bulls (Table 1): low culling rate 10% with 66 worst bulls in DGV culled, medium culling rate 50% with 328 worst bulls culled, and high culling rate 90% with 590 worst bulls culled.

Table 1. Simulated genomic pre-selection scenarios with three culling rates.

Scenario	Culling rate	No. culled bulls
Low	10%	66
Medium	50%	328
High	90%	590

For each of the three pre-selection scenarios a test run was conducted ignoring the culled bulls, and as a comparison another test run considered the culled bulls by using their generated pseudo-records. These six test runs, ignoring and considering genomic pre-selection under three levels of genomic pre-selection, were compared to the reference evaluation in order to quantify proof biases, defined as proof of a test run minus the reference evaluation. Because numbers of evaluated animals varied between evaluations, EBV from all evaluations were adjusted so that averages of common bulls with daughters were equal. In fact, the averages were nearly identical for all test runs due to the high number of common bulls between the evaluations.

3. Results

3.1. Proof biases and correlations on a population level

For all six test runs proof correlations with the reference evaluation were above 0.99 for either bulls with daughters or ancestors without data. Average proof bias for all animals was lower than 0.2% of genetic standard deviation for each of the six scenarios. Regression coefficients of proofs of the test runs on reference evaluation were nearly unity. In general, there was little problem concerning pre-selection bias appeared on the population level.

3.2. Missing animals and animals with large proof biases

Table 2 shows the number of evaluated and missing animals in the reference or test evaluations. It can be seen that more animals were missed in test runs as the culling rate increased or more bulls were culled. Using the proposed bias correction method no animals were missed for all three levels of pre-selection.

Table 2. Number of evaluated or missing animals.

Evaluation	Bulls	Ancestors	All
Reference	23557	53205	76762
Ignoring pre-selection (No. missing animals)			
Low	66	75	141
Medium	328	459	787
High	590	816	1406
Considering preselection (No. missing animals)			
All scenarios	0	0	0

Frequency of animals with proof bias equal to or greater than 10% of genetic standard deviation is shown in Table 3 for all test runs. It is obvious that the proposed bias correction method could reduce the number of animals with large proof biases significantly in all levels of culling rate. As culling rate increased, the number of animals with large proof bias was higher. The high number for test run ignoring pre-selection in case of low culling rate was possibly caused by the arbitrary definition of large bias as 10% of genetic standard deviation and the majority of them being dams without own data.

3.3. Proof bias of parents of the culled bulls

Understandably the ignorance of culled animals in conventional evaluation has more impact on proofs of their parents than remote relatives or unrelated animals. Table 4 gives proof correlation and mean, standard deviation (Std), minimum (Min) and maximum (Max) of proof bias, in percentage of genetic standard deviation, of sires of the culled genotyped bulls under all simulated scenarios. We can see that proof biases of the sires were reduced considerably for all three levels of culling rate, when the genomic pre-selection was accounted for using the proposed correction method. In addition, standard deviations and ranges of the biases went down dramatically. In all the test runs, no sire of the culled bulls was missing, indicating the sires had either own daughters or other sons not culled. Negative average proof bias from the high culling rate scenario may

indicate that the culled sons had lower genetic merit than the daughters of the sires.

Table 3. Frequency of animals with large proof bias >= 10% genetic standard deviation.

	No. bulls	No. bulls missing	No. common animals with large bias		
			bulls	ancestors	all
Test runs ignoring pre-selection					
Low	23491	66	9	6812	6821
Medium	23229	328	10	2402	2412
High	22967	590	16	7349	7365
Test runs considering pre-selection					
Low	23557	0	49	31	80
Medium	23557	0	251	202	453
High	23557	0	453	347	800

Table 4. Proof bias[§] of sires of the culled bulls.

No. sires	Correlation	Mean	Std	Min	Max
Low (ignoring / considering pre-selection)					
35	.99976	1.09	2.23	-0.28	9.60
35	.99993	0.24	1.30	-2.55	6.78
Medium (ignoring / considering pre-selection)					
80	.99931	0.39	4.10	-23.87	13.06
80	.99993	-0.13	1.34	-5.92	6.77
High (ignoring / considering pre-selection)					
101	.99917	-0.54	4.94	-21.86	11.79
101	.99991	-0.30	1.69	-8.74	6.82

[§] expressed in % genetic standard deviation

For some of those sires with all genotyped sons culled, we can see in Table 5 that average proof biases increased, in comparison to Table 4, and proof correlation dropped slightly, as expected.

When sires of the culled bulls had both culled and non-culled sons, we observed (results not shown here) that average proof biases were smaller and proof correlations slightly higher than in Tables 4 and 5. Average proof bias for high culling rate scenario was negative.

Table 5. Proof bias[§] of sires with all sons culled.

No. sires	Correlation	Mean	Std	Min	Max
Low (ignoring / considering pre-selection)					
5	.99954	2.74	3.75	0.10	9.23
5	.99981	2.00	2.83	0.07	6.78
Medium (ignoring / considering pre-selection)					
18	.99951	1.33	3.65	-3.41	11.78
18	.99990	0.49	1.79	-1.72	6.77
High (ignoring / considering pre-selection)					
27	.99905	0.08	4.86	-18.47	11.79
27	.99985	-0.05	2.13	-5.70	6.82

[§] expressed in % genetic standard deviation

Dams of the culled bulls were investigated in addition to sires. Table 6 shows proof biases of dams of the culled genotyped bulls. Note that only common dams between the test and reference evaluations were able to be analysed, 31, 166 and 317 dams of the culled bulls were missing in the test evaluation of low, medium and high culling rate, respectively.

Table 6. Proof bias[§] of dams of the culled bulls.

No. dams	Correlation	Mean	Std	Min	Max
Low (ignoring / considering pre-selection)					
31	.94979	26.11	24.98	-10.64	86.79
31	.97452	2.53	14.84	-37.80	65.34
Medium (ignoring / considering pre-selection)					
118	.95582	16.53	25.62	-58.28	88.84
118	.98468	1.49	14.44	-52.81	40.69
High (ignoring / considering pre-selection)					
172	.93363	0.95	30.98	-81.89	86.88
172	.98475	0.04	14.79	-52.42	45.47

[§] expressed in % genetic standard deviation

In comparison to the sires, proof biases of the dams were much higher and proof correlations significantly lower, which is indeed expected, because the dams had no own records, and usually one or two sons and consequently their proofs were almost completely influenced by their genotyped sons.

4. Discussion

Genomic pre-selection of young animals as parents of next generation has become a routine breeding programme worldwide. Because only genotyped young candidates with high genomic EBV will have daughters producing milk, conventional genetic evaluation based on the selected performance data will be biased. We developed a simple method to correct this pre-selection bias by generating pseudo-records for genotyped animals using DGV and associated EDC and adding them back to the conventional evaluation system. This bias correction method was validated using German Holstein phenotypic and genomic data. The proposed method has reduced proof biases significantly, particularly for dams of the culled bulls.

The proposed bias correction method utilises individual variation in DGV and reliability of genotyped animals, therefore it has been shown to be very effective in reducing proof bias. However, it can be further improved in several aspects. Firstly, generation of pseudo-record using Formula 1 was done on an animal by animal basis, a more accurate way was proposed by Ducrocq and Liu (2009). Secondly, this proposed method can be fine tuned to account for the fact that genotyped young calves share the same phenotypic information with training animals. A solution to this problem may be the one-step approach by Misztal *et al.* (2009), where all genotyped animals with or without phenotypes can be simultaneously evaluated with all non-genotyped animals in a single system. Thirdly, this study was focused on a single trait analysis, it would be more logical to use total merit index as trait to be evaluated. Finally, the simulated genomic pre-selection was based on DGV in this study, using combined genomic EBV could be more closer to reality.

The genomic pre-selection bias problem poses also a challenge for international evaluation, all culled genotyped animals must be considered in a similar way as proposed here. Because countries differ in size of genomic reference population and selection intensity of genomic pre-selection, it is even more important to solve the pre-selection problem in international genomic evaluation in order to ensure high quality of evaluation service and fair comparison among countries.

5. References

Ducrocq, V. & Liu, Z. 2009. Combining genomic and classical information in national BLUP evaluations. Interbull meeting, Barcelona, Spain. *Interbull Bulletin 40*, 172-177.

Liu, Z., Reinhardt, F., Bünger, A. & Reents, R. 2004. Derivation and calculation of approximate reliabilities and daughter yield deviations of a random regression test-day model for genetic evaluation of dairy cattle. *J. Dairy Sci. 87*, 1896-1907.

Liu, Z., Seefried, F., Reinhardt, F. & Reents, R. 2009. Dairy cattle genetic evaluation using genomic information. *Interbull Bulletin 39*, 23-28.

Meuwissen, T.H., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819-1829.

Misztal, I., Aguilar, I., Johnson, D., Legarra, A., Tsuruta, S. & Lawlor, T. 2009. A unified approach to utilise phenotypic, full pedigree, and genomic information for a genetic evaluation of Holstein final score. *Interbull meeting*, Barcelona, Spain. *Interbull Bulletin 40*, 240-243.

Patry, C. & Ducrocq, V. 2009. Bias due to selection. *Interbull Bulletin 39*, 77-82.

Reinhardt, F., Liu, Z., Seefried, F. & Thaller, G. 2009. Implementation of genomic evaluation in German Holsteins. Interbull meeting, Barcelona, Spain. *Interbull Bulletin 40*, 219-226.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91*, 4414-4423.