An optimized single-step SNP BLUP model for calf fitness in German Holstein

H. Alkhoder, Z. Liu, L. Polman, E. Pasman, J. Heise, and R. Reents
*IT-Solutions for Animal Production (**vit**), Heinrich-Schröder-Weg 1, 27283 Verden, Germany*
Corresponding author: Zengting.liu@vit.de

Abstract

A single-step SNP BLUP was introduced in routine official evaluation of German Holstein in April 2025 for all traits, including an early-measured trait, calf fitness, defined as calf survival between day 3 and 15 months after birth. Prior to the single-step model implementation, a mixed reference population of bulls and calves was set up for the calf fitness genomic evaluation using a multi-step genomic model. During the testing phase of the single-step model, an unrealistic, strong genetic trend of calf fitness was observed in genotyped animals, when compared to the multi-step genomic model or pedigree-based conventional model. Having searched for plausible causes for the overestimation, we detected a much lower mortality rate for genotyped calves than non-genotyped ones, particularly for the early periods from day 3 to 120 after birth. Although all female calves were genotyped under the whole-herd genotyping scheme in Germany, farmers did not always take genotyping sample right after the birth of a calf, causing a delay in genotyping for the early periods of the trait calf fitness. In addition, there were limited economic incentives for farmers to genotype dead calves. To solve the overprediction bias of the calf fitness evaluation, we developed a new single-step model by using only genotypes of sires of all female calves with phenotypic data. Genomic breeding values of the genotyped calves and all other genotyped animals were indirectly predicted based on SNP effect estimates and residual polygenic effect estimates of all the genotyped sires from the new single-step model. Genomic validation showed a slightly higher accuracy of the new single-step model using sire genotypes than the original model using genotypes of all animals. In comparison to a significant over-prediction for the original model, the new single-step model using only the sire genotypic records gave an almost unbiased genomic prediction. Genetic trends in genotyped AI bulls or female animals were no longer overestimated with the new single-step model. The problem of inflated genomic prediction of the original single-step model seems to be solved by using only the genotypic data of sires of female calves.

Key words: single-step model, calf fitness, genomic evaluation, prediction bias

Introduction

Calf fitness (CF) is an economically important trait for dairy farmers which was defined as female dairy calf survival from day 3 to 15 months / 458 days after birth for German dairy cattle breeds. The whole-time span was divided into five periods: days 3 to 14, 15 to 60, 61 to 120, 121 to 200, and 201 to 458, that were treated as genetically correlated traits with a multi-trait linear animal model (Heise et al. 2016).

Prior to the official implementation of a single-step SNP BLUP genomic model (SSM, Liu et al. 2014) for all evaluated traits in German Holstein in April 2025, genotype records of all animals born from 2005 onwards were used also for trait CF in the testing period of the model SSM. Thanks to the whole-herd female calf genotyping scheme introduced in 2019, more than 1 million German Holstein female calves with phenotypic CF records had also genotype data available for genomic evaluations. An unexpected overestimation of genetic trend in the genotyped animals was, however, identified for trait CF during the test phase of the SSM model. To solve the problem of overestimated genetic trend in trait CF, an alternative single-step model needed to be developed.

The objectives of this study were 1) to identify causes of the inflated genomic prediction of the single-step model using all genotype data; 2) to develop a new single-step model for removing the overestimation bias; and 3) to conduct genomic validation for the two single-step models with a full and a truncated data set.

Materials and Methods

Phenotypic, genotypic and pedigree data from August 2024 (2408) were obtained for the investigation on the trait CF. Following the Interbull GEBV test rules (Mäntysaari et al. 2010), four years of phenotype data were deleted to simulate a genomic valuation in August 2020 (2008t). Two SSM models were compared: using genotype data of all animals including all female calves and using only genotype data of sires of female calves with phenotypic records. Table 1 describes the genotype and phenotype data for the full evaluation 2408 and the truncated evaluation 2008t.

Table 1. Phenotype and genotype data for the full (2408) and truncated evaluation (2008t)

Data set	Female calves with phenotypes	Genotyped calves	Genotyped sires
2408	13,273,996	1,075,268	36,325
2008t	10,733,873	501,653	26,578
Ratio	81%	47%	73%

The total number of genotyped Holstein animals in both evaluations was 1,631,843, including 1,433,599 females and 198,244 male animals. All the genotyped animals were born in 2005 and later due to the left truncation of genotype data (Alkhoder et al. 2024). Figure 1 shows the numbers of female dairy calves with trait CF and genotyped Holstein female calves with trait CF. The solid lines represent the numbers of female dairy calves in the full evaluation 2408 with phenotypic CF records (in blue) and genotyped Holstein female calves (in orange). The dotted lines denote the numbers of female dairy calves with phenotypes in the truncated evaluation 2008t (in black) and with both phenotype and genotype data (in orange).

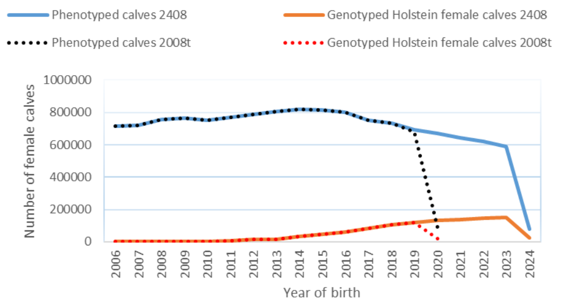

Figure 1. Numbers of female calves in the full evaluation 2408 and truncated evaluation 2008t

Figure 2 shows the number of genotyped sires of the female calves with phenotypic CF in the full evaluation 2408 (solid line) and in the truncated evaluation 2008t (dotted line).

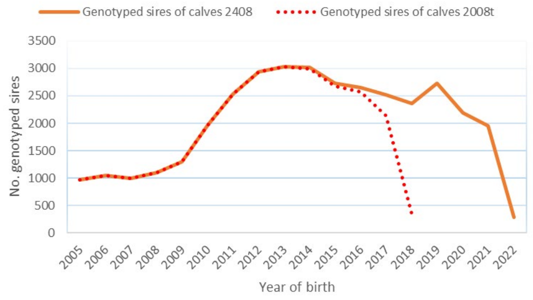

Figure 2. Numbers of genotyped Holstein sires of female calves in the full evaluation 2408 and truncated evaluation 2008t

Mortality rate of dairy female calves

Germany has run a whole-herd genotyping scheme in participating herds since 2016, where all newborn dairy female calves are to be genotyped. For legitimate reasons, farmers do not always take genotyping samples immediately after birth of a calf, causing a delay in genotyping. Furthermore, there is limited incentive for farmers to genotype dead or even sick calves. Based on all 615,927 Holstein female calves born in 2022 which had opportunity to reach the end of CF trait definition (458 days) in the evaluation of April 2025. Figure 3 shows the mortality rates of 148,427 genotyped and 467,500 non-genotyped Holstein calves with respect to the five periods of trait CF. For the first period of CF, non-genotyped calves have a mortality rate of 2.46% that is six times higher than that of genotyped calves, 0.41%. Similarly, the non-genotyped calves are 3 times more likely to die than the genotyped ones in the second period.

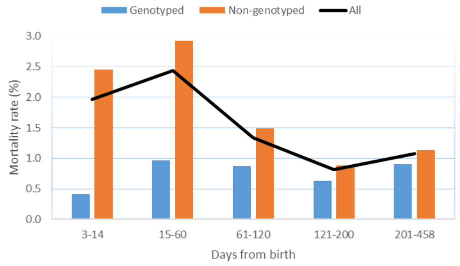

Figure 3. Mortality rates of genotyped and non-genotyped Holstein female calves

A multi-step SNP BLUP genomic model

Under the multi-step genomic model (MSM) for all other evaluated traits in German Holstein (Liu et al. 2011), a SNP BLUP model was applied to deregressed EBV of reference bulls and calves for trait CF. In the full evaluation 2408 there were 1,055,144 reference calves and 13,077 reference bulls representing their non-genotyped calves.

Two single-step SNP BLUP genomic models

For trait CF, we compared two SSM models: one using all genotype records including all calves and the other using only genotype data of sires of the female calves. A single-step SNP BLUP model (Liu et al. 2014) was applied to the phenotype data and the two genotype data sets. For the SSM with sire genotypes, GEBV of all the other genotyped animals were indirectly predicted, following the weekly genomic evaluation procedure (Alkhoder et al. 2024a).

Genotyped Holstein AI bulls and female calves

For trait CF two main groups of genotyped animals were chosen to investigate the impact of the two SSM models: genotyped Holstein AI bulls and genotyped Holstein female calves. Both animal groups were highly relevant for breeding and most affected by the SSM model change as well. Figure 4 shows the number of AI bulls by year of birth with a total of 8,391 genotyped Holstein AI bulls owned by German AI studs born from 2005 to 2023. Numbers of the genotyped Holstein female calves having trait CF born in 2010 and later are shown in Figure 5. The total number of the genotyped Holstein calves is 1,072,492 in the evaluation 2408.

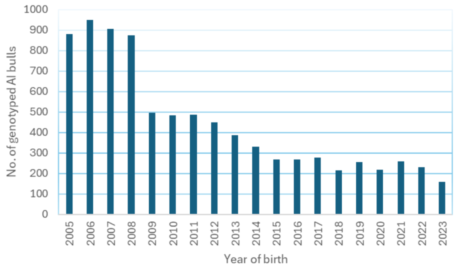

Figure 4. Number of genotyped Holstein AI bulls owned by German AI studs

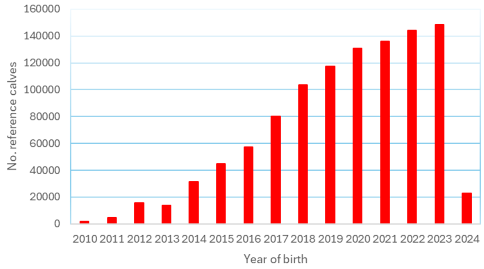

Figure 5. Number of genotyped Holstein female calves with phenotypic records

Genomic validation via GEBV Test

Following Interbull GEBV Test rules, a total of 980 validation bulls were defined based on the full and truncated data sets. The most recent version of the GEBV Test software from September 2024 was used at the time of conducting the genomic validation. To investigate the impact of dependent variable on the validation results, both deregressed GEBV and GEBV were used as target variable. For the MSM model, an earlier genomic validation with a linear regression model (Legarra and Reverter, 2018) was performed using data from a full evaluation in April 2023 (2304) and from a truncated evaluation in April 2021 (2104). For this special validation, the original SSM with genotype data of all animals was evaluated besides the MSM model. Furthermore, GEBV of 355 validation bulls from the full evaluation 2304 were regressed on GEBV from the truncated evaluation 2104 for the regression analysis.

Results & Discussion

The single-step genomic full evaluation, 2408, was run using the two genotype data sets of German Holstine for trait CF: using genotypes of all animals and using only genotypes of sires of the female calves with own phenotypic records in CF. For the GEBV test, the single-step evaluation based on the truncated phenotypic data, 2008t, was conducted for the two genotype data sets as well. In addition, we further ran the MSM model using the full data set 2408 and truncated 2008t. All the SSM evaluations were run with software MiX99 (Strandén and Lidauer, 1999), whereas our own programs were used for the MSM evaluations.

Earlier genomic validation results

Table 2 shows results of the earlier genomic validation using the linear regression method (Legarra and Reverter, 2018) by comparing the full evaluation 2304 to truncated evaluation 2104. It can be seen for both SSM and MSM models that the model R^2 value is relatively high and b_1 value close to 1. However, caution needs to be taken when interpreting the validation results, because the validation bulls have low reliability values in comparison to other traits, between 0.5 and 0.6 in the full evaluation 2304, for the low heritability trait CF. The high R^2 values may also be attributed to the lower contribution of own calves' phenotypic data to the total reliability of the validation bulls. Using GEBV as dependent variable for the linear regression may partially lead to the b_1 values close to 1, too.

Table 2. Genomic validation results using data from the full evaluation 2304 and truncated evaluation 2104

Genomic model	Number of validation bulls	R^2 value	b_1 value
Single-step	355	0.61	1.04
Multi-step	355	0.40	0.92

Genomic validation results of the two single-step models

Tables 3 and 4 give results of genomic validation for both SSM models via Interbull GEBV Test software using data from the full evaluation 2408 and truncated evaluation 2008t. The total number of validation bulls was 980. The two SSM models show significantly lower R^2 values than those in Table 2, indicating that the dependent variable deregressed GEBV results in a lower R^2 value than the dependent variable GEBV. Another explanation for the lower R^2 values is the number of years in the data truncation, 4 years for the validation in Table 3 versus 2 years for the validation in Table 2. Based on the regression slope b_1 values, we can conclude that the SSM model using genotype data of all animals failed the GEBV test, leading to overestimated candidate GEBV.

Table 3. Genomic validation results of the two single-step models using data from the full evaluation 2408 and truncated evaluation 2008t

Deregressed GEBV as dependent variable	R^2 value	b_1 value	Pass
Using only sire genotypes	0.191	0.954	PASS
Using all genotypes	0.164	0.849	FAIL

As an alternative form of dependent variable in the GEBV Test, GEBV of validation bulls from the full evaluation 2408 were regressed on those of the truncated evaluation 2008t. Table 4 shows genomic validation results of the two SSM models with GEBV as dependent variable. The R^2 values of both SSM models are nearly equal and higher than the validation using deregressed GEBV in Table 3. Based on the regression slope b_1 estimates, the two SSM models pass the GEBV test. However, the b_1 value of the SSM using all genotype data, 0.933, deviates more from its expected value of 1.

Table 4. Genomic validation results of the two single-step models using GEBV as dependent variable for the regression analysis

GEBV as dependent variable	R^2 value	b_1 value	Pass
Using only sire genotypes	0.444	0.963	PASS
Using all genotypes	0.436	0.933	PASS

GEBV of the genotyped Holstein AI bulls

Figure 6 shows genetic trends of GEBV in the genotyped Holstein bulls born between 2005 and 2023. GEBV of the AI bulls are expressed in genetic standard deviations (σ_g) in Figure 6.

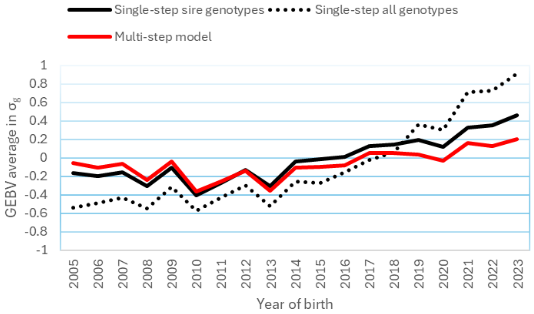

Figure 6. Genetic trends of the three genomic models in the genotyped Holstein AI bulls

The SSM model using all genotype data (dotted black line) which failed the GEBV Test (see Table 3) has too high genetic trend, with a genetic progress of 1.4 genetic standard deviations in las 10 years between 2013 and 2023, despite the fact no direct selection has been imposed on this trait CF in German Holstein. The new, optimized SSM model using only sire genotype data (solid black line) has brought down the genetic trend significantly, to a level much closer to the MSM model. For information, the genetic trend of the MSM model is equal to conventional evaluation for the genotyped AI bulls with calf phenotype data.

GEBV variances of the AI bulls by birth year are given in Figure 7. All the three genomic models have nearly equal GEBV standard deviations within the birth years.

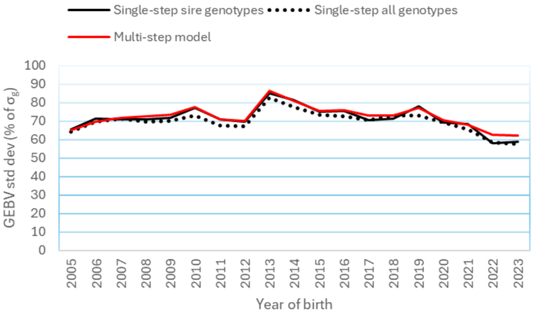

Figure 7. GEBV standard deviations of the three genomic models for the genotyped Holstein AI bulls

Figure 8 shows GEBV correlations between all pairs of three 3 genomic models. The new SSM model with only sire genotype has high GEBV correlations with either the SSM model using all genotype data (solid black line) or the MSM model (dotted green line). The MSM model and the SSM with all genotypes (dashed blue line) have the lowest GEBV correlations.

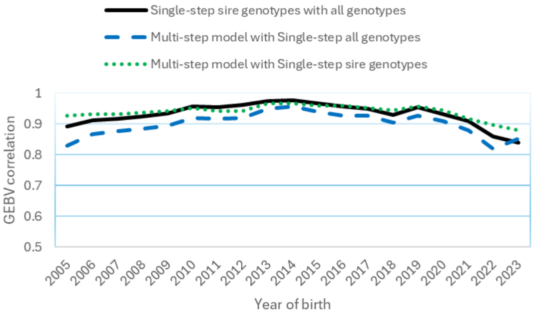

Figure 8. GEBV correlations between the genomic models for the genotyped Holstein AI bulls

GEBV of the genotyped Holstein female animals

Regarding the genetic trends of the 3 genomic models in the genotyped Holstein female calves, we can see in Figure 9 that the SSM using all genotype data (dotted black line) has severely overestimated GEBV of the female calves, due to the much lower mortality rate of genotyped than non-genotyped calves (see Figure 3). However, the unrealistically high genetic trend is reduced significantly for the SSM model when only the genotype data of sires were used (solid black line). The GEBV averages by birth year of the new SSM with only sire genotypes are now only slightly higher than those of the MSM model (solid red line).

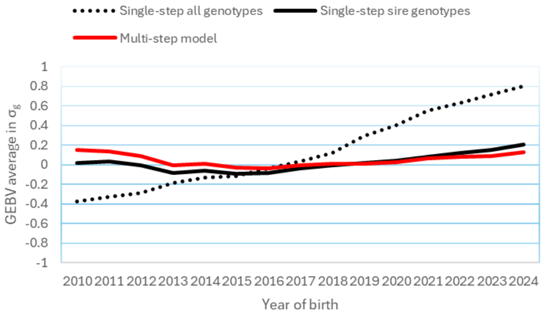

Figure 9. Genetic trends of the three genomic models in the genotyped Holstein female animals

In addition, GEBV variances in the genotyped female calves are compared among the genomic models (Figure 10). Despite the large trend difference in genotyped animals between the two SSM models, the genotyped Holstein female calves have nearly equal GEBV variances (dotted and solid black lines), probably due to the rather low heritability of trait CF. The MSM model has lower GEBV variance (solid red line) than the two SSM models, which may be explained by the contribution of non-genotyped relatives with phenotype data to the female calves GEBV of the SSM model.

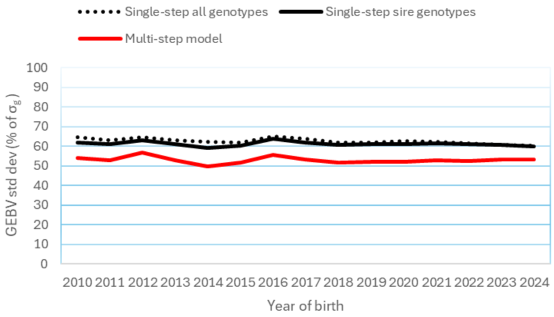

Figure 10. GEBV standard deviations of the genomic models in genotyped Holstein female calves

GEBV correlations between the genomic models are shown in Figure 11 for the genotyped Holstein female calves. The highest GEBV correlations are found between the MSM model and SSM with only sire genotypes (dotted green line), whereas the GEBV correlations between the MSM model and the SSM using all genotypes are lowest (dashed blue line).

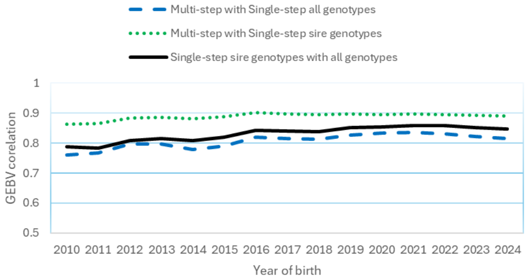

Figure 11. GEBV correlations between the genomic models for the genotyped Holstein female calves

The MSM model with a mixed reference population of bulls and calves did not show the problem of overestimated GEBV for the early measured trait CF, partly due to the reference bulls whose EBV containing phenotype data of both live and dead calves. As another contributing factor, the pseudo-phenotype data of reference bulls or calves in the MSM genomic evaluation were deregressed conventional EBV of bulls and calves, which had been estimated in the preceding conventional evaluation without consideration of any genotype data. Thus the problem of genotyped calves having a much lower mortality rate than the non-genotyped calves could not have any impact on the conventional EBV at the preceding step and on the subsequently generated deregressed conventional EBV of the reference bulls or calves.

Conclusions

The single-step genomic model using genotype data of all animals appears to give biased genomic prediction for the early-measured trait calf fitness, when genotyping of some calves are delayed with respect to the trait definition or there is a limited genotyping of dead calves. The inflated genomic prediction, observed in the single-step evaluation using all genotype data for calf fitness in German Holstein, occurred even under the whole-herd female genotyping scheme in Germany, where all female calves are systematically genotyped.

A strategy for solving the problem of inflated genomic prediction was developed by using only genotype data of sires of phenotyped female calves, because the sires have both dead and live calves and almost all the phenotyped calves have a genotyped sire. Following Interbull GEBV test rules, phenotypic data in the last four years were removed from the full evaluation to perform a genomic validation. The new, optimized single-step model resulted in a slightly higher accuracy and a nearly unbiased regression slope estimate than the original single-step model. For the low heritability trait calf fitness where validation bulls have a relatively low reliability, we found that the deregressed GEBV are clearly more appropriate as dependent variable of the linear regression than the GEBV of validation bulls. In case of a high number of reference animals for the reduced genomic evaluation, 4-year data truncation is preferred to a 2-year data cut to achieve more realistic validation results.

By comparing the new single-step model to the previous one using genotype data of all animals, we found that the genetic trends in AI bulls and genotyped female calves were reduced significantly and GEBV become more accurate, with slight change in GEBV variances. Finally, we can draw a conclusion that the inflated single-step prediction problem of the early-measured trait calf fitness has been solved by using only genotype data of calf sires.

Acknowledgments

Colleagues of Nordic Genetic Evaluation Center (NAV) are acknowledged for a video meeting on 26 September 2024 on young stock survival evaluation.

References

Alkhoder, H., Liu, Z., and Reents, R. 2024. The marker effects of a single-step random regression model for 4 test-day traits in German Holsteins. *J. Dairy Sci.* 107:410–424. https://doi.org/10.3168/jds.2023-23793.

Alkhoder, H., Liu, Z., and Reents, R. 2024a. A weekly genomic evaluation of newly genotyped selection candidates based on a single-step genomic model. *Interbull Bulletin* 60:121-126.

Heise, J., Liu, Z., Stock, K.F., Rensing, S., Reinhardt, F., and Simianer, H. 2016. The genetic structure of longevity in dairy cows. *J. Dairy Sci.* 99:1253-1265.

Legarra, A. and 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.

Liu, Z., F. R. Seefried, R. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 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.

Liu, Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. *J. Dairy Sci.* 97:5833-5850.

Mäntysaari, E., Liu, Z., and VanRaden P.M. 2010 Interbull validation test for genomic evaluations. Interbull Bulletin 41:17-22.

Strandén, I., and M. Lidauer. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. *J. Dairy Sci.* 82:2779-2787.