Słomian, D., Żukowski, K., Skarwecka, M., Ten Napel, J., Vandenplas, J., Szyda, J. (2026) Modelling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23649: https://doi.org/10.31274/wcgalp.23649 ©2026 Słomian, et al. CC BY-NC-ND

Modelling missing parents in single-step test-day SNP-BLUP evaluation of dairy cattle

D. Słomian[1*], K. Żukowski[1], M. Skarwecka[1], J. Ten Napel[2], J. Vandenplas[2], and J. Szyda[1,3]

[1]National Research Institute of Animal Production, Balice, Poland; [2]Wageningen University & Research, Wageningen, Netherlands; [3]Wroclaw University of Science, and Technology, Poland; [*]Corresponding author: dawid.slomian@iz.edu.pl

Introduction

Many countries implement a single-step model for routine evaluation (Legarra *et al*., 2014, Mäntysaari *et al*. 2017). The advantage of single-step models over currently used approaches is that they combine all available information, i.e., phenotype, genotype, and pedigree. The structure of pedigree data is an important aspect of the genetic evaluation of dairy cattle (Bradford *et al*., 2019). Nowadays, a challenge is to handle adequately missing individuals in the pedigree. The standard approach is to combine missing parents within unrelated genetic groups (Westell *et al*., 1988; Legarra *et al*., 2007). An alternative to genetic groups to deal with missing parents in the pedigree is metafounders, which involves forming groups of missing parents based on average genetic relationships from single nucleotide polymorphisms (SNPs) (Legarra *et al*., 2015). The aim of this study is to compare the approaches of handling missing data in the pedigree using the single-step SNP-BLUP model for a real dataset for fat yield in the Polish Holstein population.

Materials & Methods

The dataset (Table 1) is from the Polish national evaluation of fat yield conducted in April 2024. The phenotype data were divided into two datasets: 1) the full data set contains 63,484,231 test day records of 3,701,610 cows, and 2) the truncated data set contains 58,446,695 test day records of 3,224,577 cows. The truncated dataset corresponds to the full dataset without data for the youngest individuals, i.e., born in the last four years. The 181,999 animals have genomic information from 46,118 SNPs. Pedigree information was available for 4,712,143 individuals and was extracted up to the third generation from animals with genotypes and phenotypes using the RelaX2 software (Strandén, 2014).

Table 1. Number of animals in the analysed data set.

Data	Sex	Number of animals	Number of records
Phenotype (full data set)	Cows	3,701,610	63,484,231
Phenotype (truncated data set)	Cows	3,224,577	58,446,695
Genotype	Cows	113,019	
	Bulls	68,972	
Pedigree	Cows	4,569,044	
	Bulls	143,099	

Based on pedigree, the following scenarios were considered: 1) Pedigree_real (**P_Real**) - the original pedigree from routine evaluation containing 262,519 (5.6%) missing sires and 719,360 (15.3%) missing dams; 2) Pedigree_2010 (**P_2010**) - **P_Real** with approximately 20 % of the dams and 10% of the sires set to missing, containing 446,669 (9.5%) missing bulls and 1,076,127 (22.8%) missing cows; 3) Pedigree_4020 (**P_4020**) - **P_Real** with approximately

40% of the dams and 20% of the sires set to missing, containing 884,192 (18.7%) missing bulls and 1,868,957 (39.6%) missing cows. Differences between scenarios start with the second generation,

Moreover, three approaches for dealing with missing data in pedigree have been used: 1) **RP** - raw pedigree with missing parents' IDs set to missing; 2) **GG** - genetic groups with missing parents replaced by unrelated genetic groups, which are defined based on year of birth, country of origin, and sex; 3) **MF** - metafounders with missing parents replaced by metafounders, which represent genetic groups with relationships estimated from the genomic information of descendants.

The following single-step test-day SNP-BLUP model (Liu *et al*., 2004) was used to predict breeding values:

$$\mathbf{y}=\boldsymbol{Xh}+\boldsymbol{Wf}+\boldsymbol{Vp}+\boldsymbol{Vu}+\boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y}$ contains cow test-day records for fat yield from the first three parities, $\boldsymbol{h}$ is a vector of fixed effects of herd-test_date-parity-milking_frequency, $\boldsymbol{f}$ is a vector of fixed lactation curve coefficients, which was modelled by the Wilmink function (Liu *et al*., 2004), $\boldsymbol{p}$ is a vector of permanent environmental effects described by three random regression coefficients of the Legendre polynomial, and $\boldsymbol{u}$ is a random additive genetic effect also described by the three random regression coefficients of the Legendre polynomials. $\boldsymbol{V}$ is a design matrix for $\boldsymbol{u}$, $\boldsymbol{p}$ contains the Legendre coefficients for the first three lactations, $\boldsymbol{X}$ is the design matrix for the fixed herd-test-date-parity-milking-frequency effect $\boldsymbol{h}$, and $\boldsymbol{W}$ is the design matrix for fixed effect $\boldsymbol{f}$, and $\boldsymbol{e}$ is a residual. The model was implemented using MiXBLUP 3.0 (Vandenplas *et al*., 2022).

The validation of $\widehat{\boldsymbol{\mu}}_{\mathbf{305}}$ was conducted on $\widehat{\boldsymbol{\mu}}_{\mathbf{305}}$ representing the 305-day Genomically Enhanced Breeding Values (GEBVs) combined over all three lactations, using the following pattern:

$$GEBVf=0.5GEBV_1+0.3GEBV_2+0.2GEBV_3, \quad (2)$$

where GEBVf is the combined GEBV and GEBVi represents GEBV for the i-th lactation. Validation cows were defined as cows whose records were removed for a truncated data set; validation bulls were defined as sires born between 2017 and 2019, and having more than 20 daughters. The weights used in the calculation emphasise first lactations because genetic parameters vary across lactations; higher heritability is observed in the first lactation (Mucha *et al*., 2014). The validation test was implemented separately for cows and bulls, using linear regression:

$$GEBVf = b_0 + b_1GEBVp + \boldsymbol{e}, \quad (3)$$

where GEBVf represents the vector of 305-day GEBVs predicted based on the full data set, while GEBVp represents GEBVs predicted based on the truncated data set, b_0 represents the intercept, which indicates a systematic bias in the model's prediction, and b_1 represents the regression slope, the dispersion of prediction compared to actual results. The R^2 coefficient is obtained from the linear regression (3) and serves as a measure of prediction accuracy. It indicates the percentage of variance in the GEBpVp explained by GEBVf.

Results & Discussion

Validation. Validation results are reported for 562 validation bulls (387 genotyped and 175 ungenotyped) and 482,810 validation cows (30,227 genotyped and 452,336 ungenotyped). Fig. 1A shows the estimated slopes (b_1) of the validation models divided by scenario, sex, parity, and genotyping status. For bulls, the value of b_1 was close to the expected value of one for most scenarios, except **P_2010** (1.271) and **P_4020** (1.328) for **MF** for ungenotyped bulls. Moreover, for **MF**, overdispersion was observed when comparing the predictions of **P_2010** and **P_4020** with **P_Real**. Fig. 1B shows intercepts (b_0) for all scenarios. Similar estimates, close to zero, were observed for every scenario, where the minimum value was -0.463 (**P_2010**

for **RP** for genotyped bulls) and the maximum value was 0.067 (**P_4020** for **MF** for genotyped cows). In general, the intercepts estimated for bulls were negative, whereas for cows, the values were very close to zero. Fig. 1C shows R^2, and Fig. 1D displays the Pearson correlation coefficients between GEBVs predicted from the truncated and full datasets for the validation animals, divided by scenario, sex, parity, and genotyping status. When **MF** or **GG** were used, the R^2 and correlations for cows were generally higher than those for the bulls. Additionally, the GEBV of genotyped cows always resulted in higher correlations than those of ungenotyped cows, regardless of missing parent handling. However, for **P_Real**, R^2 and correlations depended on the adopted missing parent approach. The most striking scenario was for the ungenotyped **P_Real**, where the correlation for **RP** was 0.892, decreased to 0.870 for **GG**, but increased to 0.924 for **MF**. The correlations for bulls followed a different pattern. For ungenotyped bulls, the correlation in each case increased sequentially from **RP** through **GG** to **MF**. For genotyped bulls, for **P_Real**, the correlation was similar across scenarios, but for **P_2010** and **P_4020**, **MF** resulted in slightly lower correlations.

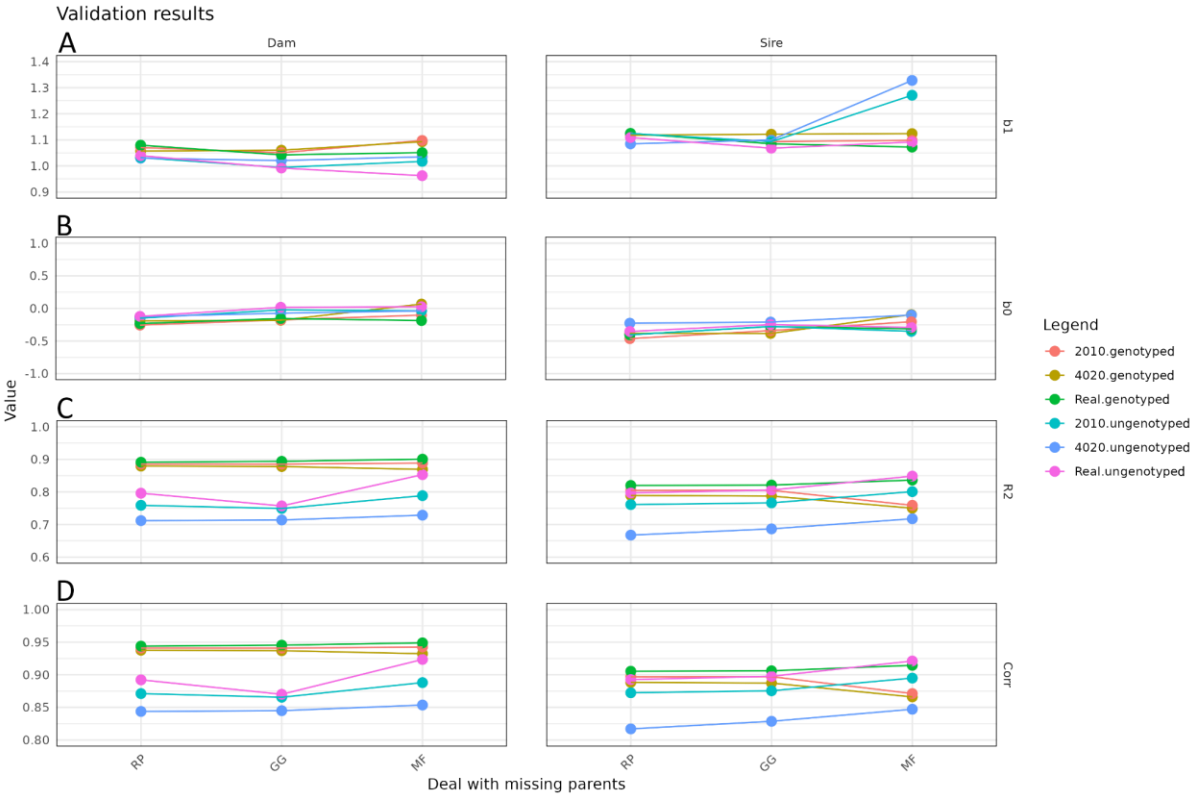

Figure 1. Validation results for individuals divided by sex and method. Section A presents regression slopes (b_1), Section B presents intercepts (b0), Section C presents coefficients of determination (R2), and Section D presents Pearson correlations (Corr).

GEBV comparison. Fig. 2 shows the average GEBV trends for all scenarios divided by sex. The average GEBV trend is a combination of the population average (genetic trend: year-to-year change in the mean breeding value) and the intensity of selection (the advantage of selected parents within the year). After 2010, we observed that the mean GEBV increased for all animals, especially bulls. For each scenario, the highest increase in the GEBV trend was realized by **P_Real** with missing parents handled via **MF.**

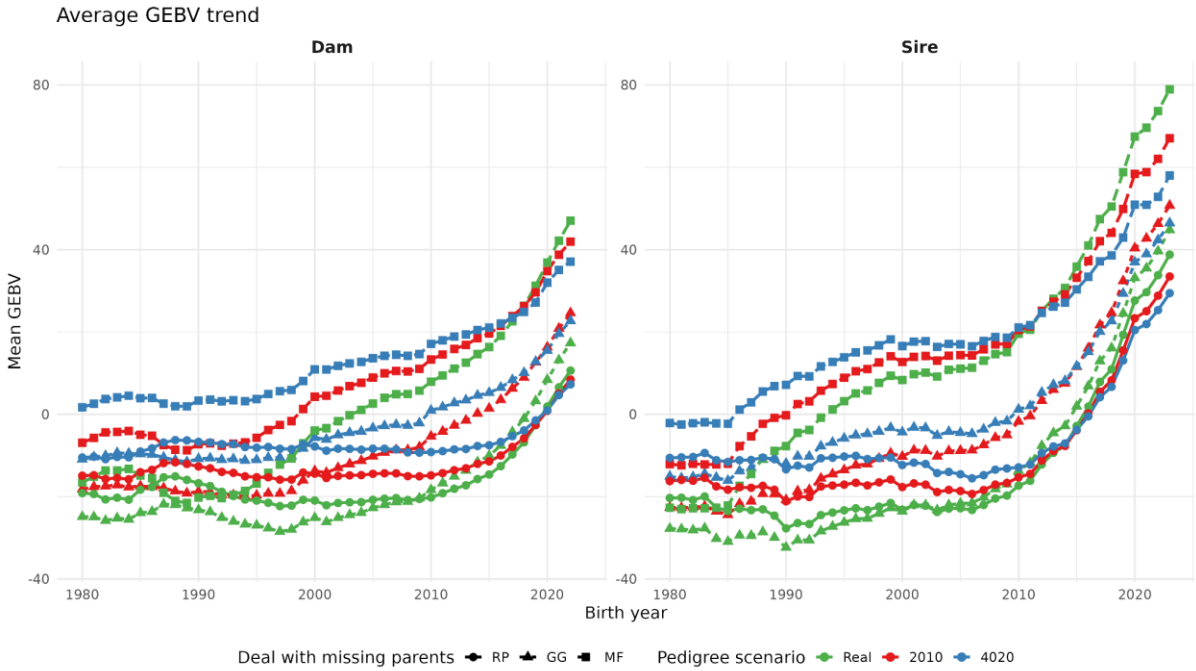

Figure 2. Genetic trends for individuals divided by sex and method.

Conclusions

This study demonstrates that methods for handling missing parents in pedigrees may impact GEBV and that handling missing parents is increasingly important with the increasing number of incomplete pedigrees. The most important result of this study is that using the metafounder approach may lead to biased predictions for ungenotyped individuals, particularly as the proportion of missing parents increases. This bias arises because, as the number of metafounders increases, the model becomes less effective at transmitting genomic information to ungenotyped individuals. Finally, their breeding values are based on misidentified relationships, leading to increased bias due to the pedigree's incompleteness. In contrast, for genotyped individuals, no marked differences in the handling of missing parent data were observed.

References

Legarra A., Bertrand J., Strabel T., *et al*. (2007) J. Anim. Breed. Genet, 124:286–295. https://doi.org/10.1111/j.1439-0388.2007.00671.x

Legarra A., Christensen O. F., Aguilar I., *et al*. (2014) Livest. Sci., 166, 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra A., Christensen O. F., Vitezica Z. G., *et al*. (2015) Genetics 200:455–468. https://doi.org/10.1534/genetics.115.177014

Liu Z., Reinhardt F., Bünger A., *et al*. (2004) J. Dairy Sci, 87:1896–1907. https://doi.org/10.3168/jds.s0022-0302(04)73348-2

Mäntysaari E. A., Evans R. D., and Strandén I. (2017) J. Anim. Sci., 95:4728–4737. https://doi.org/10.2527/jas2017.1912

Mucha S., Mrode R., Coffey M., *et al*. (2014) J. Dairy Sci. 97:2455–2461. https://doi.org/10.3168/jds.2013-7319

Strandén I. (2014) RelaX2 program for pedigree Analysis, User's Guide for Version 1.65.

Vandenplas J., Veerkamp R., Calus M., *et al*. (2022) MiXBLUP 3.0 – software for large genomic evaluations in animal breeding programs. Proceedings of 12th World Congress on Genetics Applied to Livestock Production, 1498–1501. https://doi.org/10.3920/978-90-8686-940-4_358