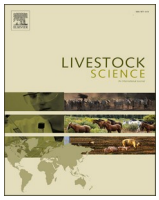
Livestock Science 306 (2026) 105919

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci



Influence of phenotyping truncation on genetic parameter estimate and genomic prediction of productive traits in nellore cattle

Gabriel Gubiani [a,*], Larissa Temp [b], Miller Teodoro [b], Daniel Cardona-Cifuentes [c], Marisol Londoño-Gil [b], Letícia Pereira [d], Roney Teixeira [b], Claudio Ulhôa Magnabosco [e], José Bento Ferraz [a], Fernando Baldi [a]

[a] Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of São Paulo, Pirassununga, Brazil
[b] Department of Animal Science, São Paulo State University – Júlio de Mesquita Filho (UNESP), Prof. Paulo Donato Castelane, Jaboticabal, Brazil
[c] Facultad de Ciencias Agrarias, Fundación Universitaria Agraria de Colombia - UNIAGRARIA, 111166, Bogotá, Colombia
[d] Association of Breeders and Researchers (ANCP), Ribeirão Preto, Brazil
[e] Brazilian Agricultural Research Corporation (Embrapa Cerrados), Planaltina, Federal District, Brazil

HIGHLIGHTS

- We evaluated how phenotypic data truncation affects genomic prediction in Nellore cattle.
- Strategic removal of historical data maintains accuracy, bias, and dispersion while reducing computational demands.
- Excluding recent phenotypes (retrospective scenario) sharply reduces accuracy, heritability, and genetic variance.
- Low-heritability traits with sparse records, such as residual feed intake, are highly sensitive to data loss.

ARTICLE INFO

Keywords:
Bos indicus
Genomic selection
Phenotypic truncation

ABSTRACT

This study evaluated the impact of progressively reducing phenotypic data collection on variance components and the genomic prediction ability of productive traits in Nellore cattle. A total of 288,456 phenotypic records and 51,471 genotypes were obtained from the National Association of Breeders and Researchers (ANCP), covering three traits: rib eye area (REA), residual feed intake (RFI), and adjusted weight at 450 days (W450). Two data-reduction scenarios were simulated: a retrospective scenario, involving the sequential removal of the most recent data by year, and a prospective scenario, involving the annual exclusion of historical data. In both scenarios, animals born in 2021 comprised the validation set. Analyses were performed using the single-step genomic best linear unbiased prediction method (ssGBLUP), with prior estimation of variance components adjusted for each scenario and year. Prediction accuracy, bias, and dispersion were evaluated using linear regression (LR) between genomic estimated breeding values (GEBV) from the complete and reduced datasets. In the retrospective scenario, starting from the complete 2021 database and removing contemporary records down to 2013, prediction accuracy declined from 0.72 to 0.299 for REA, from 0.51 to 0.11 for RFI, and from 0.77 to 0.34 for W450, accompanied by reduced heritability estimates and increased underdispersion. These effects were mainly attributed to the loss of genomic connectivity and the low proportion of genotyped animals (<1% in recent years). For RFI, genetic variance also increased due to the concentration of data on farms with strong genomic structure, although the absence of contemporary records limited predictive ability. Conversely, in the prospective scenario, starting from 2013 dataset and progressively removing the recent records through 2021, accuracy remained essentially stable, changing from 0.72 to 0.73 for REA, from 0.51 to 0.50 for RFI, and from 0.76 to 0.77 for W450, while bias and dispersion stayed within acceptable limits. This outcome was supported by the presence of up-to-date phenotypes and a higher proportion of genotyped animals (exceeding 70% by 2018). These findings demonstrate that while the exclusion of historical data can be a viable strategy to reduce computational demands, maintaining a contemporary and representative phenotypic dataset is critical to ensuring the reliability of genomic predictions and the long-term success of breeding programs.

* Corresponding author.
E-mail address: gabriel.gubiani@usp.br (G. Gubiani).

https://doi.org/10.1016/j.livsci.2026.105919
Received 11 September 2025; Received in revised form 25 January 2026; Accepted 10 February 2026
Available online 12 February 2026

1. Introduction

In recent years, research in genetic improvement has increasingly focused on genomic annotation and phenotyping (Steibel, 2023). Despite major advances in molecular and reproductive technologies, genetic progress still largely depends on the availability of phenotypic variability for complex traits and the analysis of limited production data, which informs selection and mating decisions. Accurate phenotyping, particularly for traits that are difficult to measure or require large-scale data, remains a challenge in animal breeding. Compared to genotyping, phenotyping is typically more expensive, labor-intensive, and susceptible to errors (Ventura et al., 2020).

Before the advent of genomics, estimating the effects of Mendelian sampling was difficult due to the limited availability of individual phenotypic data (VanRaden and Wiggans, 1991). Robust phenotypic information is essential for models to adequately capture the variation arising from genetic sampling. Expanding the phenotypic database enhances the predictive ability of models by increasing the diversity of information and enabling a more accurate estimation of genetic variance among individuals (Goddard and Hayes, 2009).

Almeida et al. (2016) evaluated the impact of progressively reducing phenotypic data collection on milk production traits in Holstein cattle. The authors reported a decline in prediction accuracy and an increasing tendency toward bias as phenotypic information decreased. Similarly, Vitezica et al. (2011) demonstrated that incorporating diverse phenotypic data, such as slaughter records and ultrasound measurements, improved the accuracy of genetic predictions and increased genetic variability in swine. Daetwyler et al. (2014) also found that including additional data, such as slaughter weight, subcutaneous fat thickness, and final meat pH, enhanced the genomic prediction of meat quality traits in beef cattle. These findings suggest the importance of integrating phenotypic data from various sources and stages of the production cycle, as the diversity and continuous updating of such information enhance the assessment of genetic variability and improve model robustness.

However, some studies indicate that phenotypic data from earlier generations have a limited effect on the accuracy of genomic predictions for young animals from more recent generations. Steyn et al. (2023) reported that although older bulls still contribute some genetic variability to current populations, their influence on prediction accuracy is lower than that of younger bulls in the Holstein breed. Similarly, Lourenco et al. (2014) found that information from previous generations had minimal impact on the accuracy of genomic predictions for young genotyped animals. The authors suggest that using data from the last two or three generations is sufficient to maintain prediction accuracy, primarily because of their closer genetic relationship to the validation population and the sharing of larger genomic segments.

In beef cattle, traits related to carcass composition and meat quality, such as fat thickness, rib eye area (REA), and marbling, are associated with productivity and profitability (Al-Jammas et al., 2016). However, accurate evaluation of these traits requires specific techniques, including ultrasound scanning and carcass analysis (Muller, 1987). Similarly, traits associated with feed efficiency, such as residual feed intake (RFI) (Koch et al., 1963), are essential for identifying genetically superior animals capable of reducing feed requirements and minimizing environmental impact (Sainz et al., 2024). Recording RFI demands specialized equipment and trained personnel, such as the GrowSafe® or Intergado® systems, which substantially increase the cost of phenotypic data collection. For instance, measuring RFI involves expensive feed efficiency trials to monitor daily dry matter intake (DMI), while adjusted weight at 450 days (W450) requires specific management protocols and precise weighing schedules.

With genomic selection and the ability to select young genotyped animals without phenotypic data, breeders have shown increasing interest in reducing data collection for traits that are difficult to measure, without compromising prediction accuracy. This interest is driven by the high cost of phenotyping traits such as carcass composition and feed efficiency, as well as the operational and economic benefits provided by genomic selection. However, concerns remain about the effects of discontinuing phenotypic data collection, particularly regarding prediction accuracy and the evaluation of new or complex traits. The aim of this study was to simulate the impact of progressively reducing phenotypic data collection, taking into account the use of records from both young and historical animals, on variance components and the genomic prediction ability for feed efficiency, growth, and carcass composition traits in Nellore cattle.

2. Materials and methods

2.1. Ethics statement

This study was exempt from evaluation by the Animal Ethics Committee (CEUA), as established by Law No 11,794 of 08/10/2008 and Normative Resolution No 51 of 05/19/2021 from the National Council for Animal Experimentation Control (CONCEA) because the data was obtained from an existing database.

2.2. Phenotypic data

The phenotypic data used in this study were provided by the Nellore Brazil animal breeding program, coordinated by the National Association of Breeders and Researchers (ANCP, Ribeirão Preto, SP, Brazil). Records were collected from animals born between 2013 and 2021, originating from the Midwest, Southeast, North, and Northeast regions of Brazil. The dataset comprised a total of 288,456 phenotypic records, including 121,023 records for REA, 17,049 for RFI, and 150,384 for W450.

The herds were mainly managed under extensive production systems typical of tropical livestock farming, with pastures dominated by *Urochloa brizantha* cultivar. Feeding strategies varied across farms and included protein and mineral supplementation, particularly during the dry season, as well as the use of nitrogen sources such as urea. On some farms, calves received creep feeding. Mating predominantly took place between November and January. The pedigree file contained 315,097 records, including 13,044 bulls and 170,101 dams.

2.3. Genomic data

The database used in this study was provided by the National Association of Breeders and Researchers (ANCP) and included 51,471 Nellore cattle genotyped using the CLARIFIDE® Nellore 4.0 panel, which contains approximately 70,000 SNP markers. Imputation accuracy (from the 30 K LD panel to the 77 K HD panel) and the percentage of correctly imputed genotypes were 98.5 % and 97.9 %, respectively, as reported by Carvalheiro et al. (2014). Quality control of the genotypic data was performed with the PREGSF90 software (Aguilar et al., 2014). SNP samples with a call rate below 0.90 were removed. Additionally, to verify kinship, markers with Mendelian conflicts exceeding 1 %, allele frequencies below 0.05, and those located on non-autosomal chromosomes were excluded.

2.4. Traits

Phenotypic traits were obtained through direct measurements and specific analyses. Residual feed intake was calculated as described by Silva Neto et al. (2023), based on the difference between observed and estimated intake adjusted for metabolic body weight and average daily gain. The trait W450 was determined from individual weighing conducted on animals between 405 and 495 days of age. Rib eye area was measured by ultrasound in animals aged 15 to 18 months. Images were captured between the 12th and 13th ribs of the *Longissimus dorsi* muscle using a 3.5-megahertz linear transducer coupled to a B-mode scanner (Aloka SSD 500, Tokyo, Japan), following the guidelines of the

Ultrasound Guidelines Council and the methodology of the Beef Improvement Federation (Hohenboken, 2002).

Contemporary groups (CG) were defined based on farm, year, season of birth, and management group for the traits W450 and REA. For RFI, the efficiency trial was also included in the CG definition. Contemporary groups with fewer than four records were excluded from the analysis, as were phenotypic records exceeding ±3 standard deviations from the CG mean. Additionally, the age of the animal at the start of the trial was included as a covariate in the model for RFI.

2.5. Scenarios

Two scenarios were evaluated to investigate the impact of progressively excluding either recent or historical phenotypic data on variance components and genomic predictions. For both scenarios, animals born in 2021 were used exclusively as a validation set, meaning their phenotypic records were removed from the prediction analyses. The training dataset was modified each year according to the truncation criterion of each scenario (retrospective or prospective), allowing the evaluation of how different reference set compositions influence variance components and genomic predictions for young animals. In both scenarios, a specific reference year was defined as corresponding to the complete dataset (whole), from which annual temporal truncation was subsequently applied.

In the retrospective scenario (RS), the year 2020 was defined as the reference year corresponding to the complete dataset (whole), the phenotypic data from the most recent cohorts were sequentially excluded, starting with animals born in 2020 and moving backward to 2013. At each annual truncation, phenotypic records of younger animals were removed from the dataset, retaining only individuals from older generations. The objective was to evaluate the impact of missing recent phenotypic data on the variance components and genomic predictions for young animals born in 2021.

In the prospective scenario (PS), the year 2013 was defined as the reference year corresponding to the complete dataset (whole), the phenotypic data from older cohorts were excluded sequentially, starting with animals born in 2013 and progressing to 2020. At each annual truncation, the oldest records were removed, preserving only the most recent phenotypic information. The objective was to assess the impact of excluding historical data on variance components, genomic predictions, and the accuracy of genomic prediction for young animals born in 2021. The number of animals with phenotypes, genotypes associated with phenotypes, and their proportion for each scenario and each trait are shown in Table 1.

2.6. Genetic parameters estimation

The analyses were performed using multi-trait mixed animal models based on pedigree information to estimate genetic parameters for REA, RFI, and W450 for each combination of year and phenotypic truncation scenario. The objective was to obtain specific values for additive genetic variance (σ^2_a) and residual variance (σ^2_e), adjusted to the data structure of each evaluated scenario. The (co)variances were estimated using the blupf90+ software, part of the BLUPF90 program family (Misztal et al., 2014). The model used were as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_u} + \boldsymbol{e} \quad (1)$$

where $\mathbf{y}$ is the vector of phenotypic records, $\boldsymbol{\beta}$ is the vector of fixed effects (CG and dam age at calving (DAC)) and linear and quadratic effects of animal age as covariates, $\mathbf{X}$ is the incidence matrix associating $\boldsymbol{\beta}$ with $\mathbf{y}$, $\boldsymbol{u}$ is the vector of direct additive genetic effects, $\boldsymbol{Z}$ is the incidence matrix associating $\boldsymbol{u}$ with $\mathbf{y}$, and $\mathbf{e}$ is the residual effect.

All random effects were assumed to follow multivariate normal (MVN) distributions. Specifically, the vector of additive genetic effects and the vector of residuals were modeled as follows:

$$u \sim MVN(0, A \otimes V), \cdot e \sim MVN(0, I \otimes R) \quad (2)$$

where $\boldsymbol{A}$ is the pedigree-based relationship matrix, $\mathbf{I}$ is an identity matrix of appropriate dimensions, and $\mathbf{V}$ and $\mathbf{R}$ are the additive genetic and residual (co)variance matrices, respectively, defined as:

$$V = \begin{bmatrix} \sigma^2_{u_1} & \sigma_{u_{12}} & \sigma_{u_{13}} \\ \sigma_{u_{21}} & \sigma^2_{u_2} & \sigma_{u_{23}} \\ \sigma_{u_{21}} & \sigma_{u_{22}} & \sigma^2_{u_3} \end{bmatrix}; R = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} & \sigma_{e_{13}} \\ \sigma_{e_{21}} & \sigma^2_{e_2} & \sigma_{e_{23}} \\ \sigma_{e_{31}} & \sigma_{e_{32}} & \sigma^2_{e_3} \end{bmatrix};$$

Table 1
Descriptive statistics, including number of records (N), mean, standard deviation, and number of genotyped animals, for W450, RFI, and REA across retrospective and prospective scenarios from 2013 to 2021.

	Trait		2021	2020	2019	2018	2017	2016	2015	2014	2013
Retrospective	**W450**	**N**	149.324	127.548	108.081	90.424	75.837	63.173	51.213	40.934	31.092
		Mean	319.270	315.270	310.360	305.630	302.570	300.440	298.150	295.380	292.56
		SD	58.108	56.820	55.397	54.131	54.352	54.340	55.664	56.442	57.348
		Genotypes	43.481	29.874	19.762	13.349	9.033	5.952	3.441	960	130
	RFI	**N**	15.430	12.112	7.914	5.297	3.023	2.268	1.809	1.476	1.046
		Mean	−0.00096	−0.00077	−0.00162	−0.00201	−0.00196	−0.00086	−0.00007	0.00089	−0.00507
		SD	0.732	0.722	0.702	0.713	0.746	0.714	0.655	0.644	0.64164
		Genotypes	10.790	7.995	4.691	2.655	1.126	0.750	0.629	0.583	307
	REA	**N**	118.919	94.622	75.235	60.423	47.914	37.801	29.702	22.915	16.526
		Mean	59.484	58.690	57.050	56.870	56.010	55.190	54.800	54.440	54.385
		SD	11.802	11.881	11.313	11.276	11.347	11.497	11.598	11.657	11.955
		Genotypes	48.800	30.506	16.883	9.052	4.466	2.140	0.174	0.138	103
Prospective	**W450**	**N**	149.324	21.401	40.800	58.230	72.797	85.379	97.277	107.494	117.316
		Mean	319.200	342.820	342.770	340.450	336.660	333.230	330.430	328.390	326.37
		SD	58.194	60.111	58.580	57.818	56.857	56.938	56.326	56.228	56.326
		Genotypes	43.483	13.607	23.719	30.132	34.448	37.529	40.040	41.922	43.352
	RFI	**N**	15.430	3.318	7.516	10.133	12.407	13.162	13.621	13.954	14.384
		Mean	−0.00096	−0.00165	−0.00027	−0.00041	−0.00072	−0.00098	−0.00108	−0.0012	−0.00099
		SD	0.732	0.769	0.762	0.742	0.729	0.735	0.742	0.741	0.73833
		Genotypes	10.790	7.795	6.099	8.135	9.664	10.040	10.161	10.207	10.483
	REA	**N**	118.919	24.297	43.684	58.496	71.005	81.118	89.217	96.004	102.393
		Mean	59.484	62.565	63.680	62.183	61.827	61.485	61.045	60.689	60.307
		SD	11.802	10.957	11.441	11.725	11.522	11.402	11.451	11.513	11.568
		Genotypes	48.800	18.294	31.917	39.751	44.334	46.660	48.626	48.662	48.697

W450, adjusted weight at 450 days; RFI, residual feed intake; REA, ribeye area. The retrospective scenario corresponds to the progressive exclusion of recent phenotypic information, while the prospective scenario corresponds to the sequential exclusion of older records.

The conditional distribution of the phenotypes is:

$$y|\beta, u \sim N(X\beta + Zu, R),$$

A is the relationship matrix, **I** is the identity matrix, and $\sigma^2_{u_i}$, $\left(\sigma_{u_{ij}}\right)$, σ^2_{ei}, $\left(\sigma_{eij}\right)$, where $i, j = 1$ for REA, 2 for RFI and 3 for W450, represent the respective direct additive genetic and residual (co)variances. An inverse scaled Wishart distribution with non-informative hyperparameters was used as the prior for all (co)variance matrices, and a flat (non-informative) prior was assumed for the fixed effects vector $\boldsymbol{\beta}$.

The resulting estimates were then used in the genetic value prediction stage, ensuring that each year was analyzed using its corresponding variance parameters, consistent with the level of phenotypic truncation applied.

2.7. Genomic prediction

The GEBV were predicted using the single-step genomic best linear unbiased prediction method (ssGBLUP) (Aguilar et al., 2010), implemented with the BLUP90IOD3 software (Lourenço et al., 2022; Misztal et al., 2014), based on the same model as Eq. (1). However, the assumptions considered for the genomic prediction model were:

$$\begin{bmatrix} u \\ \mathrm{e} \end{bmatrix} \sim \mathrm{N}\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{H}\sigma^2_u & 0 \\ 0 & \boldsymbol{I}\sigma^2_e \end{bmatrix}\right], \tag{3}$$

Where **I** is an identity matrix with dimensions equal to the number of animals with records, and **H** represents the covariance structure among animals in the ssGBLUP framework (Legarra et al., 2009), with its inverse defined as described by Aguilar et al. (2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}^{-1}_{22} \end{bmatrix}, \tag{4}$$

where $\mathbf{A}^{-1}_{22}$ is the inverse of the pedigree-based relationship matrix for genotyped animals, and $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix. Estimates of variance components without incorporating genomic data were obtained using the BLUPF90+ program, part of the BLUPF90 software suite (Misztal et al., 2014; Lourenco et al., 2022).

2.8. Validation

The animals used for validation were selected from those born in 2021, with complete genotype and phenotype information and no progeny records. The final validation set included 1423, 1487, and 1035 animals for REA, RFI, and W450, respectively. The same set of validation animals was used across all analyses conducted between 2013 and 2020. Two datasets were used to evaluate the impact of phenotype removal: whole dataset, which included both genotypic and phenotypic data from the validation animals and was used to estimate the total GEBV ($\widehat{\boldsymbol{u}_w}$), and partial dataset, which for the validation animals included only genotypic data, excluding their phenotypes, to calculate the partial GEBV ($\widehat{\boldsymbol{u}_p}$).

The whole dataset represents the biological benchmark, as it includes the phenotypes of the 2021 validation animals to estimate the true population parameters. In contrast, the baseline for genomic prediction corresponds to a validation setting in which the complete historical dataset (2013–2020) is retained, but the phenotypes of the 2021 cohort are masked to simulate a real validation setting.

This validation strategy simulates a practical scenario in genomic selection, where the genomic estimated breeding values of young animals, that is, selection candidates, are predicted using phenotypic and genotypic information from proven or older animals in the population.

The impact of phenotype removal on genomic predictions was then evaluated using the linear regression (LR) method, followed the framework proposed by Legarra and Reverter (2018), and focused on comparing accuracy, bias, and dispersion parameters. The accuracy of genomic prediction was calculated as the square root of the partial reliability of the GEBV:

$$\widehat{\mathrm{acc}} = \sqrt{\frac{\mathrm{cov}\left(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p\right)}{(1-\overline{\mathrm{F}})\widehat{\sigma^2_u}}}, \tag{5}$$

where $\widehat{\boldsymbol{u}_w}$ represents the whole GEBV, and $\widehat{\boldsymbol{u}_p}$ represent the partial GEBV, $\overline{\mathrm{F}}$ is the estimated mean inbreeding coefficient for animals included in the validation set, and $\boldsymbol{\sigma}^2_{\boldsymbol{u}}$ is the additive genetic variance. The bias was estimated from the average genetic values:

$$\mu w_P = \frac{\overline{\hat{\boldsymbol{u}}}\boldsymbol{p} - \overline{\hat{\boldsymbol{u}}}\boldsymbol{w}}{\sigma_u}, \tag{6}$$

The dispersion was assessed by regressing the whole GEBV on the partial dataset of GEBV:

$$b_{w,p} = \frac{cov\left(\hat{\mathrm{u}}_w, \hat{\mathrm{u}}_p\right)}{var\left(\hat{\mathrm{u}}_p\right)}, \tag{7}$$

where the regression coefficient $b_{w,p}$ is expected to have a value of $E\left(b_{w,p}\right) = 1$ in the absence of overdispersion or underdispersion. Values below 1 indicate inflation or overdispersion, while values above 1 suggest deflation or underdispersion. According to Tsuruta et al. (2011), deviations of up to 5 percent from the expected value are typically considered accurate, whereas deviations of up to 15 percent are still regarded as acceptable.

3. Results and discussion

The present study investigated how the progressive reduction of phenotypic data influences variance components, as well as the accuracy, bias, and dispersion of genomic predictions. The analysis of variance components revealed distinct patterns between the retrospective and prospective scenarios, emphasizing the influence of the temporal structure of the data on the estimation of additive genetic variance, residual variance, and heritability, as presented in Table 2.

In the RS, a consistent reduction in genetic variance was observed for traits such as REA and W450 as the most recent data were removed. As noted by Dekkers (1992), additive genetic variance and the accuracy of predicted genetic values can be affected by both selection and data availability. In this case, the decrease in additive variance was more pronounced than the reduction in residual variance, leading to successive declines in heritability. This pattern can be explained by the reduced amount of information available for estimating genetic parameters, which can compromise accuracy, particularly when the data structure loses depth and connectivity. However, for the RFI trait, the behavior was the opposite: additive variance increased with the exclusion of the most recent data. In the early years of data collection, records originated mainly from research institutions, such as universities and experimental centers, involving a small number of farms with strong genetic connectivity among animals and more controlled environmental conditions. This composition of the initial database favored the estimation of high genetic variances and higher heritability, even with less diversity among herds. As the database expanded to include information from different herds, environmental and residual variability increased, resulting in lower estimates of genetic parameters.

In the PS, the behavior of the variance components was more erratic. For REA and W450, genetic variance remained relatively stable in the early stages of excluding older data but showed an initial upward trend, reaching a peak around 2019, before stabilizing at a slightly lower level in 2020 This indicates that the older data had little influence on the estimation of components. In contrast, the more recent records, which were more numerous and had a higher proportion of genotyped and phenotyped animals, sustained higher levels of additive variance and,

Table 2
Estimated variance components and heritability in retrospective and prospective scenarios for three productive traits in Nellore cattle.

Traits[2]	Year	Retrospective[1]			Prospective[1]		
		σ^2a	σ^2e	h^2	σ^2a	σ^2e	h^2
REA	whole	13.75	25.65	0.35	13.75	25.65	0.35
	2021	13.71	25.68	0.35	13.71	25.68	0.35
	2020	13.10	25.94	0.34	17.50	23.48	0.43
	2019	11.92	25.90	0.31	18.65	23.87	0.44
	2018	11.60	25.26	0.31	17.62	24.77	0.42
	2017	11.01	24.53	0.31	16.00	26.03	0.38
	2016	11.27	24.29	0.32	15.45	25.81	0.37
	2015	10.40	24.43	0.30	14.93	25.94	0.36
	2014	10.48	24.54	0.30	14.52	25.85	0.36
	2013	9.65	25.56	0.27	14.45	25.61	0.36
RFI	whole	0.09	0.45	0.17	0.09	0.45	0.17
	2021	0.10	0.46	0.17	0.10	0.46	0.17
	2020	0.09	0.45	0.17	0.09	0.52	0.15
	2019	0.11	0.41	0.21	0.08	0.51	0.14
	2018	0.11	0.42	0.21	0.09	0.48	0.15
	2017	0.13	0.45	0.23	0.09	0.46	0.16
	2016	0.13	0.41	0.25	0.08	0.48	0.14
	2015	0.14	0.32	0.31	0.08	0.48	0.15
	2014	0.15	0.29	0.34	0.08	0.48	0.15
	2013	0.12	0.31	0.29	0.09	0.47	0.16
W450	whole	301.96	546.25	0.36	301.96	546.25	0.36
	2021	303.85	542.11	0.36	303.71	543.26	0.36
	2020	295.96	526.84	0.36	359.22	568.49	0.39
	2019	290.69	522.93	0.36	422.42	516.11	0.45
	2018	280.71	515.96	0.35	408.76	506.86	0.45
	2017	278.29	516.01	0.35	386.85	507.76	0.43
	2016	256.68	536.2	0.32	386.85	507.76	0.43
	2015	242.79	547.7	0.31	364.26	515.75	0.41
	2014	239.68	545.92	0.30	347.36	524.88	0.40
	2013	254.93	542.97	0.32	333.19	528.53	0.39

σ^2_a, additive variance; σ^2e, residual variance; h^2, heritability; REA, rib eye area; RFI, residual feed intake; W450, adjusted weight at 450 days of age.
Note: Whole = benchmark dataset including phenotypes of validation animals (2021). 2021 = baseline prediction scenario in which phenotypes of 2021 animals were masked for validation.

consequently, higher heritability. For RFI, the pattern again contrasted with that observed for carcass and weight traits. The removal of older data resulted in a less structured and less connected dataset, as feed efficiency records are limited to a small number of farms. This loss of structure compromised the consistency of the estimates. This resulted in inconsistent estimates of genetic variance, reinforcing that the impact of truncation is strongly influenced by the original composition of the database and the distribution of the trait across herds.

These findings emphasize that the temporal pattern of phenotypic data collection has a direct impact on the estimation of variance components, with both the representativeness and the distribution of records playing a critical role in the robustness of genetic parameter estimates (Pszczola et al., 2012). Additionally, as noted by Weng et al. (2016), trait heritability and the genetic connectedness between training and validation populations are essential for enhancing predictive ability. Traits with lower heritability demand a larger number of generations with consistent records to achieve reliable estimates and are therefore more vulnerable to the removal of historical data (Mrode, 2014; Hollifield et al., 2021). In this sense, costly and difficult-to-measure traits such as RFI are more prone to fluctuations in genetic parameter estimates, as their records are typically concentrated in a limited number of farms. Maintaining a representative and continuously updated database under these conditions is essential to ensure the robustness of genomic predictions and the reliability of selection decisions.

Obtaining phenotypic information in beef cattle still presents logistical and financial challenges, including the number of observations required, the need for skilled labor, and the variability of collection conditions (Akdemir and Isidro-Sánchez, 2019). Although advances in genotyping have reduced costs, phenotyping remains a major expense (Misztal et al., 2020), prompting questions about the necessity of continuous phenotype collection in genomic selection programs. Fig. 1 summarizes the results for REA, RFI, and W450, enabling the observation of accuracy, bias, and dispersion trends of GEBV under progressive phenotypic data removal in both retrospective and prospective scenarios.

In RS, starting from the complete 2021 database and removing recent years down to 2013, prediction accuracy declined from 0.72 to 0.30 for REA, from 0.51 to 0.11 for RFI, and from 0.77 to 0.34 for W450. These losses were driven not only by the removal of contemporary records but also by the reduction in the number of observations and the proportion of animals with both genotypic and phenotypic information. For REA and W450, the decline in accuracy was further compounded by reductions in genetic variance over time, lowering heritability and predictive ability. In contrast, RFI exhibited an increase in genetic variance due to the concentration of data on farms with strong genomic structure, although predictive ability was still limited by the absence of recent data. These findings underscore the importance of maintaining a comprehensive and up-to-date database to ensure strong connections between reference animals and young candidates for selection. Similar conclusions were drawn by Lourenco et al. (2014) and Cardona-Cifuentes et al. (2024), who emphasized that historical data often contribute less to the genetic evaluations of young animals, particularly after the adoption of genomics in breeding programs.

In PS, prediction accuracy remained virtually unchanged when removing historical records from 2013 through 2021, with accuracy varying only slightly from 0.72 to 0.73 for REA, from 0.51 to 0.50 for RFI, and from 0.76 to 0.77 for W450. This stability indicates that the availability of recent phenotypes was sufficient to sustain robust genomic predictions, even with the exclusion of older data. Maintaining updated phenotypic data strengthens the genomic connection between individuals in the reference and validation sets, thereby enhancing predictive ability. This link between connectivity and prediction accuracy has been widely documented. Gondro et al. (2013) and Zhang et al. (2018) emphasize that the reliability of GEBV depends on the strength of the genetic relationships between training animals and selection candidates. When this connection weakens, as observed in the RS, the ability of genomic models to capture true genetic relationships is reduced.

Bias also exhibited marked changes in the RS, moving from a near-zero bias (0.00) to −0.77 for REA, from −0.01 to 0.10 for RFI, and from −0.02 to −0.67 for W450, indicating a strong miscalibration when recent records were excluded. Studies such as Bussiman et al. (2023) also highlight that excessive reliance on outdated data can increase bias and undermine prediction quality, suggesting that the careful exclusion of such data may be an effective strategy.

This genetic disconnect between candidates and the reference population reduces the model's ability to capture linkage disequilibrium between markers and quantitative trait loci (QTL), thereby compromising predictive ability (Calus et al., 2008; Daetwyler et al., 2008; Goddard, 2009). These differences can be attributed to the genetic architecture of each trait. RFI, with its lower heritability and greater environmental influence, depends more heavily on a representative and updated phenotypic database. In contrast, traits such as W450 and, to a lesser extent, REA, which have higher heritability and phenotypic representation, show greater predictive ability even as contemporary data are reduced.

In PS, bias showed only slight variations, remaining close to zero (0.00) for REA, ranging from −0.01 to 0.02 for RFI, and from −0.02 to 0.01 for W450. These patterns indicate greater stability in predictions when the most recent data were retained. These findings underscore the importance of monitoring bias as a critical aspect of predictive accuracy, particularly for traits sensitive to the quality and timeliness of phenotypic records.

The dispersion of genomic predictions was also influenced by data truncation. In RS, dispersion decreased from 1.010 to 0.921 for REA,

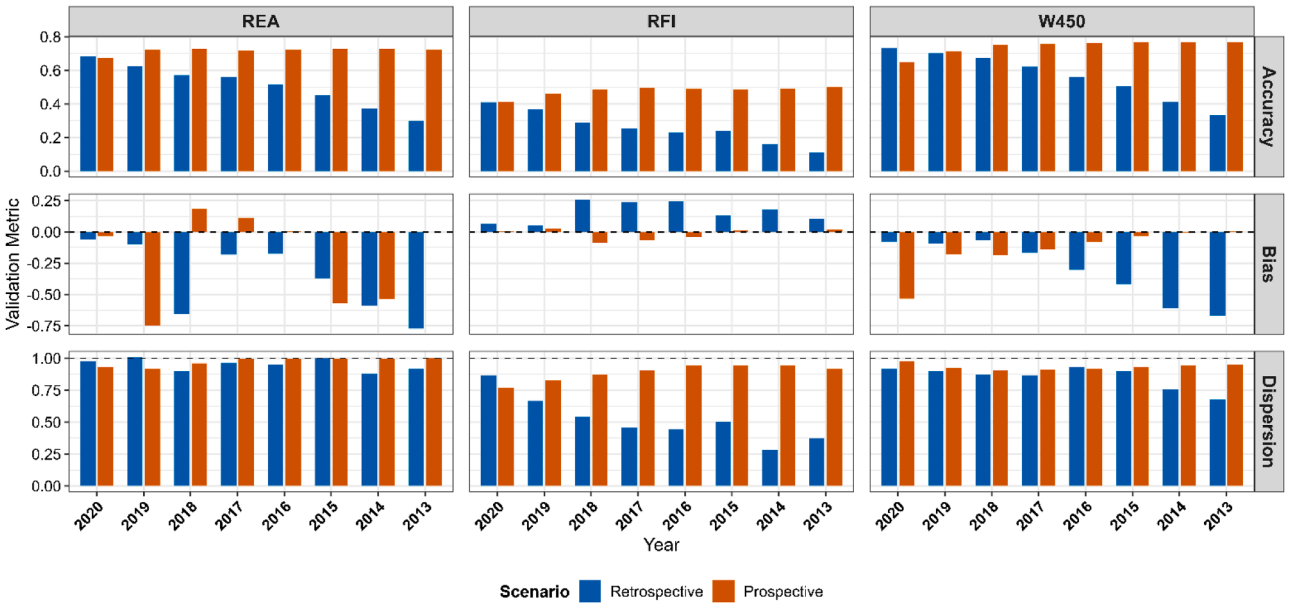


Fig. 1. Genomic prediction ability for rib eye area (REA), residual feed intake (RFI), and adjusted weight at 450 days of age (W450) under sequential truncation of phenotypic data in two scenarios: retrospective (removing the most recent records) and prospective (removing the oldest records). Prediction ability was assessed using the linear regression approach. **The X-axis indicates the year marking each training dataset; the 2021 cohort is not shown because it corresponds to the independent validation set.**

from 0.888 to 0.372 for RFI, and from 0.961 to 0.679 for W450 when recent years were excluded, indicating progressive underdispersion and reduced reliability. In PS, dispersion remained near the target, showing no change for REA (1.010 in both 2013 and 2020), increasing from 0.888 to 0.921 for RFI, and decreasing slightly from 0.962 to 0.952 for W450. This suggests that the presence of contemporary phenotypes supports better calibration of genomic models. According to Tsuruta et al. (2011), the ideal dispersion value is close to 1, with deviations up to 5 % considered acceptable. Therefore, the dispersion values obtained in the PS fall within the expected range, while those in the RS indicate reduced reliability in predictive estimates.

The proportion of genotyped animals among those phenotyped (Table 1) helps to explain the variations observed in the model's predictive ability. In RS, removing the most recent data drastically reduced this proportion, falling below 1 % for REA by 2015. This low genotypic representativeness weakened the connection between the reference and validation populations and negatively affected prediction quality. Studies such as Wolc et al. (2011) emphasize that genomic connectivity is essential for reliable estimates. In PS, retaining contemporary data maintained high proportions of genotyped animals, above 70 % until 2018 for most traits, which supported the robustness of the predictions.

These results also support previous findings on the truncation of phenotypic and pedigree data. Granado-Tajada and Ugarte (2024), in dairy sheep, and Stein et al. (2023), in cattle, demonstrated that strategically removing old data can enhance the accuracy of evaluations. Furthermore, Howard et al. (2018), Misztal et al. (2014), and Aguilar et al. (2011) showed that this approach can reduce the computational costs of genomic models, particularly in populations with large numbers of genotypes, such as the Nellore breed in Brazil.

Therefore, while excluding historical data may improve model efficiency, maintaining updated phenotypes and a high proportion of genotyped animals is essential to ensure robust genomic predictions. As emphasized by Meuwissen et al. (2001), Calus (2010), and Wolc et al. (2011), continuously updating the reference population is critical to prevent declines in prediction reliability across generations. Continuous phenotyping remains a strategic practice to keep models aligned with the current genetic landscape, thereby supporting the accuracy, stability, and effectiveness of predictions over time (Pszczola and Calus, 2016).

4. Conclusion

The removal of historical phenotypic data proved to be an effective strategy for optimizing computational resources, as long as the database remains current and representative. Data from older generations contributed little to the prediction of young animals, particularly when genomic connectivity was low. In the retrospective scenario, excluding contemporary data impaired accuracy, with a greater effect on high heritability traits such as REA and W450, due to reduced genetic variance and underdispersion of predictions. For RFI, heritability increased despite the decline in accuracy, reflecting the concentration of records in a limited number of farms. In the prospective scenario, removing older data preserved accuracy, bias, and prediction calibration, indicating that careful truncation can be safely applied in populations with strong genotypic representation, ongoing phenotyping, and updated pedigree information.

Author statement

I, **Gabriel Gubiani**, corresponding author, declare that this manuscript is original, has not been published previously, and is not currently being considered for publication elsewhere. I confirm that the manuscript has been read and approved by all listed authors, that all authors meet the ICMJE authorship criteria, and that there are no other individuals who meet these criteria but are not included as authors.

I also confirm that all authors agreed on the order in which they appear in the manuscript, that we comply with the journal's policies, including ethical guidelines and data-sharing requirements, and that there are no financial or non-financial conflicts of interest to declare.

CRediT authorship contribution statement

Gabriel Gubiani: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Larissa Temp:** Writing – original draft, Visualization, Conceptualization.

Miller Teodoro: Writing – original draft, Validation, Methodology, Formal analysis. **Daniel Cardona-Cifuentes:** Writing – original draft, Visualization, Validation, Methodology, Conceptualization. **Marisol Londoño-Gil:** Validation, Methodology, Conceptualization. **Letícia Pereira:** Data curation. **Roney Teixeira:** Conceptualization. **Claudio Ulhôa Magnabosco:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Data curation. **José Bento Ferraz:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Fernando Baldi:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no conflicts of interest related to this work.

Acknowledgements

The authors are grateful to the National Association of Breeders and Researchers (ANCP) for providing phenotypic, genomic, and pedigree data. Financial support from the São Paulo Research Foundation (FAPESP) through Grant No. 2023/03659-5 is also acknowledged.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93 (2), 743–752. https://doi.org/10.3168/jds.2009-2730.

Aguilar, I., Misztal, I., Legarra, A., Tsuruta, S., 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128 (6), 422–428. https://doi.org/10.1111/j.1439-0388.2010.00912.x.

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A., Wang, H., 2014, August. PREGSF90–POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production, 3. WCGALP, Vancouver, BC, Canada.

Al-Jammas, M., Agabriel, J., Vernet, J., Ortigues-Marty, I., 2016. The chemical composition of carcasses can be predicted from proxy traits in finishing male beef cattle: a meta-analysis. Meat. Sci. 119, 174–184. https://doi.org/10.1016/j.meatsci.2016.04.029.

Bussiman, F., Chen, C.Y., Holl, J., Bermann, M., Legarra, A., Misztal, I., Lourenco, D., 2023. Boundaries for genotype, phenotype, and pedigree truncation in genomic evaluations in pigs. J. Anim. Sci. 101 (7). https://doi.org/10.1093/jas/skad273 skad273.

Calus, M.P.L., 2010. Genomic breeding value prediction: methods and procedures. Animal 4 (2), 157–164. https://doi.org/10.1017/S1751731109991352.

Calus, M.P.L., Meuwissen, T.H.E., De Roos, A.P.W., Veerkamp, R.F., 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178 (1), 553–561. https://doi.org/10.1534/genetics.107.080838.

Carvalheiro, R., Boison, S.A., Neves, H.H., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y. T., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46 (1), 69. https://doi.org/10.1186/s12711-014-0069-1.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., Hayes, B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46 (8), 858–865. https://doi.org/10.1038/ng.3034.

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10 (6), 381–391. https://doi.org/10.1038/nrg2575.

Granado-Tajada, I., Ugarte, E., 2024. Impact of truncating historical data on prediction ability of dairy sheep selection candidates. Animal 18 (8), 101245. https://doi.org/10.1016/j.animal.2024.101245.

Hollifield, M.K., Lourenco, D., Bermann, M., Howard, J.T., Misztal, I., 2021. Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations. J. Anim. Sci. 99 (4), skab085. https://doi.org/10.1093/jas/skab085.

Koch, R.M., Swiger, L.A., Chambers, D., Gregory, K.E., 1963. Efficiency of feed use in beef cattle. J. Anim. Sci. 22 (2), 486–494. https://doi.org/10.2527/jas1963.222486x.

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 (1), 53. https://doi.org/10.1186/s12711-018-0426-6.

Lourenco, D.A.L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T.J., Forni, S., Weller, J.I., 2014. Are evaluations on young genotyped animals benefiting from the past generations? J. Dairy Sci. 97 (6), 3930–3942. https://doi.org/10.3168/jds.2013-7769.

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. https://doi.org/10.1093/genetics/157.4.1819.

Misztal, I., Legarra, A., Aguilar, I., 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97 (6), 3943–3952. https://doi.org/10.3168/jds.2013-7752.

Silva Neto, J.B., Mota, L.F., Amorim, S.T., Peripolli, E., Brito, L.F., Magnabosco, C.U., Baldi, F., 2023. Genotype-by-environment interactions for feed efficiency traits in Nellore cattle based on bi-trait reaction norm models. Genet. Sel. Evol. 55 (1), 93. https://doi.org/10.1186/s12711-023-00867-2.

Pszczola, M., Calus, M.P.L., 2016. Updating the reference population to achieve constant genomic prediction reliability across generations. Animal 10 (7), 1018–1024. https://doi.org/10.1017/S1751731115002785.

Sainz, R.D., Baldi, F., Temp, L.B., Ribeiro, L.B., 2024. Estimation of genetic parameters for maintenance energy requirements and residual feed intake in Nellore cattle. J. Anim. Breed. Genet. 141 (6), 643–655. https://doi.org/10.1111/jbg.12870.

Steibel, J.P., 2023. 431 Awardee talk: phenomics and phenotyping in animal breeding. J. Anim. Sci. 101 (Suppl. 3), 17. https://doi.org/10.1093/jas/skad281.021.

Steyn, Y., Lawlor, T.J., Lourenco, D., Misztal, I., 2023. The importance of historically popular sires on the accuracy of genomic predictions of young animals in the US Holstein population. JDS Commun. 4 (4), 260–264. https://doi.org/10.3168/jdsc.2022-0299.

VanRaden, P.M., Wiggans, G.R., 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74 (8), 2737–2746. https://doi.org/10.3168/jds.S0022-0302(91)78453-1.

Ventura, R.V., e Silva, F.F., Yáñez, J.M., Brito, L.F., 2020. Opportunities and challenges of phenomics applied to livestock and aquaculture breeding in South America. Anim. Front. 10 (2), 45–52. https://doi.org/10.1093/af/vfaa008.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93 (5), 357–366. https://doi.org/10.1017/S001667231100022X.

Weng, Z., Wolc, A., Shen, X., Fernando, R.L., Dekkers, J.C., Arango, J., Garrick, D.J., 2016. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. Genet. Sel. Evol. 48 (1), 22. https://doi.org/10.1186/s12711-016-0198-9.

Wolc, A., Arango, J., Settar, P., Fulton, J.E., O'Sullivan, N.P., Preisinger, R., Habier, D., Fernando, R.L., Garrick, D.J., Dekkers, J.C.M., 2011. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genet. Select. Evol. 43 (1), 23. https://doi.org/10.1186/1297-9686-43-23.