AIMS Animal Science, 1 (1): 51–64.
DOI: 10.3934/aas.2025005
Received: 13 May 2025
Revised: 28 August 2025
Accepted: 01 September 2025
Published: 08 September 2025

https://www.aimspress.com/journal/aas

Research article

The practical detection and correction for the test-day milk yields records of dairy cows

Jian Yang[1,3], Xiuxin Zhao[3], Xiao Wang[1], Lingling Wang[3], Guanghui Xue[3], Yan Liu[3], Zhaowei Yuan[3], Fen Pei[3], Xiaoman Li[3], Xueyan Lin[2,*], Yundong Gao [3,*] and Jianbin Li [1,*]

[1] Institute of Animal Science and Veterinary Medicine, Shandong Academy of Agricultural Sciences, Jinan, China
[2] College of Animal Science and Veterinary Medicine, Shandong Agricultural University, Tai'an, China
[3] Shandong OX Livestock Breeding Industry Co., Ltd, Jinan, China

* **Correspondence:** Email: linxueyan@sdau.edu.cn, gaoyundong@sdox.cn, msdljb@163.com.

Abstract: In this study, we applied a detection and correction approach to identify and adjust abnormal test-day milk yield records of Holstein dairy cows from the Dairy Herd Improvement (DHI) database in Shandong Province, China. By dividing the lactation period into ascending and descending phases, we used adjacent test-day data to predict expected yields and correct deviations through multiple regression analysis. The correction improved lactation curve fitting and 305-day milk yield calculations, which was crucial for generating accurate and consistent data. These improvements enhanced the reliability of genetic evaluations and herd management decisions by providing a clearer representation of milk production patterns. Following correction, data variability was reduced, as reflected by a decreased coefficient of variation. Additionally, the lactation curves more accurately captured the natural progression of milk yield, from gradual increase to peak and subsequent decline. The study highlights the importance of precise data correction for maintaining high-quality DHI records, ultimately supporting better genetic and production assessments in dairy cattle.

Keywords: DHI; Abnormal milk yield; evaluation

1. Introduction

Dairy herd improvement (DHI) has become increasingly essential as it is closely associated with genetic evaluations [1]. Data from DHI constitute the core foundation for the estimation of sire breeding values. Therefore, ensuring the accuracy of data in DHI is crucial. In practical production, some individual DHI records for Test-Day Milk Yields (TDMY) are abnormally high or low, deviating significantly from typical lactation patterns [2,3]. These anomalies compromise the accuracy of data reporting and the reliability of research and applications based on TDMY records. Studies have shown that the accuracy of raw data directly impacts the reliability and precision of genetic evaluations [4]. Madsen et al. [5] demonstrated that the presence of outliers in genetic evaluations could result in aberrant breeding values for certain cows. Researchers employ two major approaches to handle such anomalous data: One is to exclude the data, the other is to replace it with population averages, but both of which lead to loss of information. As DHI programs continue to expand, the development of more accurate and practical methods for handling anomalous records could play a critical role in enhancing the quality of data.

Gao et al. [4] proposed a computationally efficient method for detecting abnormal TDMY using an approximation of the Mahalanobis distance, which has been applied in Nordic Holstein and Red Dairy cattle populations. While effective in identifying outliers, this method focuses solely on removing such records without implementing any correction or adjustment procedures. This approach may lead to the loss of valuable information and potentially reduce the representativeness of the dataset. In contrast, more recent methods aim not only to detect anomalies but also to appropriately adjust them, thereby enhancing model robustness and maximizing data utilization. According to the guidelines provided by the Dairy Cattle Data Center of China, any TDMY less than 2 kg or greater than 90 kg is considered abnormal and should not be recorded as TDMY. Additionally, data showing significant deviations from adjacent test days should also be considered as potentially abnormal. Thus, a method for identifying abnormal TDMY records based on comparisons with adjacent test-day yields is required.

In addition to milk yield records, the development of precision livestock farming (PLF) technologies has enabled real-time monitoring of behavioral traits such as rumination and lying time, which are useful for detecting health or reproductive issues [6,7]. These advances highlight the growing importance of integrating diverse data types while emphasizing the need for accurate and corrected milk yield data as a foundation for reliable genetic evaluations. In this study, data from Chinese Holstein cows were used to develop a multiple linear regression model for predicting TDMY based on adjacent test-day records. The predicted values were compared with the original TDMY data for evaluation and correction of the original test-day records. The accuracy of this method in identifying and correcting abnormal TDMY was also assessed.

2. Materials and methods

2.1. Data source

The data used in this study were obtained from the DHI database in Shandong Province, China, including 524,512 Holstein cows from 346 dairy farms between 2004 and 2023. A total of 8,328,402 DHI test-day records were initially collected, of which 3,733,845 records remained after five quality control steps (Table 1). This set of data is designated as dataset A, which were used for estimating regression coefficients on adjacent TDMY. To validate the accuracy of the proposed method for data evaluation and correction, 2,965,909 test-day records from the DHI database were evaluated and corrected (Table 2). This set of data is designated as dataset B. It should be noted that stricter data quality criteria were applied when estimating regression coefficients to ensure model accuracy, while slightly more relaxed thresholds were used during data correction to enable broader data inclusion.

Table 1. Quality control process of dataset A.

Step	Item	Criterion	Data size
0	Raw Data	—	8,048,135
1	TDMY	≥ 18.1，≤ 57.6	6,157,793
2	DIM	≥ 5，≤ 330	5,530,807
3	Number of Test-Day Records within Lactation	≥ 3	5,303,980
4	Parity	≤ 3	4,474,035
5	Test Interval	≥ 25，≤ 39	3,733,845

Table 2. Quality control process of dataset B.

Step	Item	Criterion	Data size
0	Raw Data	—	8,328,402
1	DIM	≥ 5，≤ 330	6,680,182
2	Number of Test-Day Records within Lactation	≥ 3	6,473,832
3	Test Interval	≥ 25，≤ 55	2,965,909

2.2. Prediction for TDMY

Following Wiggans et al. [8], the lactation period was divided into two phases, ascending and descending, based on days in milk (DIM). Considering that the peak lactation typically occurs around 50 days post-calving, DIM values of 50 days or less were defined as the ascending phase, while values greater than 50 days were classified as the descending phase. For each lactation phase, a set of regression coefficients was estimated to fit the gradually increasing and gradually decreasing TDMY during these respective phases. The estimation of regression coefficients was performed according to Equation 1.

$$\begin{aligned}(y_i - y_{i-1})/(DIM_i - DIM_{i-1}) = b_0 + b_1 DIM_{i-1} + b_2 DIM_{i-1}^2 \\ + b_3 y_{i-1} + b_4 (DIM_{i-1}) y_{i-1} + e,\end{aligned} \quad (1)$$

In the equation, y_i represents the milk yield on the i-th test day; y_{i-1} denotes the milk yield on the (i-1)-th test day; DIM_i indicates the days in milk on the i-th test day; and DIM_{i-1} signifies the days in milk on the (i-1)-th test day. The term $(DIM_{i-1})y_{i-1}$ represents the interaction between the days in milk on the (i−1)-th test day and the milk yield on the (i-1)-th test day. The coefficients b0 to b4 denote the regression coefficients, while e represents the random error in the regression.

Due to the lack of previous TDMY records for the first test day, subsequent TDMY were utilized to estimate a separate set of regression coefficients for predicting TDMY on the first test day [8].

Using the estimated regression coefficients $b_{0\ldots4}$ and the data from the subsequent test day records, the slope of the change in TDMY $\widehat{b_1}$ for the first test day was calculated according to Equation 2. Similarly, using the records from the previous test day, the slope of the change in TDMY $\widehat{b_i}$ for the second and subsequent test days was computed based on Equation 3.

$$\widehat{b_1} = b_0 + b_1 DIM_2 + b_2 DIM_2^2 + b_3 y_2 + b_4 (DIM_2) y_2, \tag{2}$$

In the equations, $\widehat{b_1}$ represents the slope of the change in TDMY for the first test day; b0 to b4 are the regression coefficients estimated for different lactation phases; DIM_2 indicates the days in milk for the second test day; and y_2 is the milk yield for the second test day.

$$\widehat{b_i} = b_0 + b_1 DIM_{i-1} + b_2 DIM_{i-1}^2 + b_3 y_{i-1} + b_4 (DIM_{i-1}) y_{i-1}, \tag{3}$$

In the equations, $\widehat{b_i}$ denotes the slope of the change in TDMY for the i-th test day; b0 to b4 are the regression coefficients estimated for different lactation phases; DIM_{i-1} signifies the days in milk for the (i−1)-th test day; and y_{i-1} is the milk yield for the (i−1) -th test day.

Milk yield for the first test day was predicted using Equation 4.

$$\widehat{y_1} = y_2 + \hat{b}(DIM_2 - DIM_1), \tag{4}$$

In the equation, $\widehat{y_1}$ represents the predicted TDMY for the first test day within a parity; y_2 is the original TDMY for the second test-day within the parity; and $DIM_2 - DIM_1$ indicates the interval between the first and second test days.

Milk yield for the subsequent test days was predicted using Equation 5.

$$\widehat{y_i} = y_{i-1} + \hat{b}(DIM_i - DIM_{i-1}), \tag{5}$$

In the equation, $\widehat{y_i}$ denotes the predicted TDMY for the i-th test day within the same parity; y_{i-1} is the original TDMY for the (i−1)-th test day within the parity; and $DIM_i - DIM_{i-1}$ signifies the interval between adjacent test days.

2.3. Correcting test day milk yield records

To ensure the accuracy and reliability of TDMY data, we performed systematic correction and adjustment based on the natural lactation curve of dairy cows. Specifically, for each test-day record, we predicted the expected milk yield using data from adjacent test days, reflecting the typical pattern where milk yield rises to a peak after calving and gradually declines.

Following the International Committee for Animal Recording (ICAR) guidelines [9], we identified anomalous records as those where the observed milk yield deviated substantially from the predicted value; if the observed yield was less than 60% or greater than 150% of the predicted yield. Instead of discarding these records, we adjusted them by setting values below 60% to 60% of the predicted yield and capping those above 150% at 150% of the predicted yield. This method preserved valuable data while reducing the influence of extreme or erroneous observations, thus improving the overall data quality for subsequent analyses.

2.4. Evaluation methods for milk yield prediction accuracy

To evaluate the accuracy of TDMY prediction and the effectiveness of data correction, a multiple linear regression model was first established using Dataset A. Dataset B, which included cows with abnormal test-day records, was subsequently assessed and corrected. The data in Dataset B was stratified by parity and the number of abnormal observations. Following correction, two validation subsets, Validation Set A and Validation Set B, were randomly selected from the adjusted Dataset B. Validation Set A comprised cows with at least ten test-day records per parity and no abnormal data (a total of 507,352 test-day records), whereas Validation Set B included cows with abnormal records and was further categorized by parity and the number of anomalies (a total of 438,973 test-day records). The accuracy of the predicted TDMY was evaluated using regression analysis and Pearson correlation coefficients between the observed and predicted values.

In particular, to assess the predictive performance more precisely, a regression analysis was conducted using 507,352 test-day records from Validation Set A. The observed milk yield was treated as the independent variable (X-axis), and the predicted yield as the dependent variable (Y-axis). Scatter plots were generated separately for first-parity cows and for cows in their second and later parities. The slope of the fitted regression line and the coefficient of determination (R^2) were calculated to quantify the agreement between observed and predicted values.

To further evaluate the effectiveness of the correction method, the mean, standard deviation, and coefficient of variation of both daily milk yield and 305-day milk yield were compared before and after correction using t-tests. In addition, lactation curves were fitted at both the individual and population levels using the Wood model [10], which estimated key parameters (a, b, c) and derived secondary traits such as persistency, peak yield, and days to peak. Quality control procedures were implemented to ensure the reliability of the fitted curves. A stage-specific analysis of lactation curve variations was also conducted to investigate the effects of data correction on lactation curve parameters.

The test-day records of Validation Set B were divided into ten lactation stages, starting from the 5th day postpartum and then grouped into intervals of 30 days. For each parity group (first parity and second or higher parity), the average TDMY within each lactation stage was calculated. These averages were used to fit the population-level lactation curves for the two parity groups, respectively. Subsequently, the fitted parameters of the lactation curves before and after correction were compared to evaluate the effects of data adjustment on the lactation pattern.

2.5. Investigation and analysis of daily milk yield data

After evaluating and correcting the daily milk yield test-day records, we assessed the data quality by calculating the proportion of abnormal records across different years, parities, and days in milk (DIM). The distribution patterns of abnormal test-day records were analyzed to identify potential systematic issues, and the possible causes for the occurrence of such anomalies were further investigated.

3. Results

3.1. Regression coefficients from the multiple linear regression model

The estimates of regression coefficients from the multiple linear regression model were based on adjacent TDMY. Regression coefficients for the periods before and after peak lactation were estimated separately for the first test day and for subsequent test days. Table 3 presents the four sets of estimated regression coefficients (b0 to b4) obtained from Dataset A, where $\boldsymbol{b_0}$ represents the intercept, and b1 to b4 correspond to slopes associated with adjacent test-day yields. These coefficients reflect the varying dynamics of milk production across different lactation stages (B1 to B4), capturing the increasing trend before peak lactation and the gradual decline afterward. The distinction between coefficients for the first and subsequent test days underscores the importance of stage-specific modeling to accurately predict TDMY.

Table 3. Estimated regression coefficients for different lactation stages.

Factors	B1	P-value	B2	P-value	B3	P-value	B4	P-value
b_0	3.21×10^{-2}	< 0.001	-1.73×10^{-1}	< 0.001	7.29×10^{-1}	< 0.001	2.84×10^{-1}	< 0.001
b_1	-1.50×10^{-2}	< 0.001	-8.69×10^{-4}	< 0.001	-1.57×10^{-2}	< 0.001	-2.26×10^{-4}	< 0.001
b_2	1.69×10^{-4}	< 0.001	1.88×10^{-6}	< 0.001	1.43×10^{-4}	< 0.001	1.17×10^{-7}	< 0.001
b_3	2.88×10^{-2}	< 0.001	5.19×10^{-3}	< 0.001	-1.04×10^{-2}	< 0.001	-7.29×10^{-3}	< 0.001
b_4	-3.07×10^{-4}	< 0.001	6.77×10^{-6}	< 0.001	4.91×10^{-5}	< 0.001	-7.19×10^{-6}	< 0.001

Note: *All p-values are < 0.001, indicating highly significant differences*; $b0\ to\ b4$ denotes the regression coefficient; B1 and B2 represent the predictions for periods before and after the peak lactation, excluding the first test day; and B3 and B4 represent the predictions for periods before and after the peak lactation for the first test day.

3.2. Evaluation of the accuracy of predicted TDMY

Regression analyses were performed on Validation Set A, which contained 507,352 records. For first-parity cows, the regression slope was 0.66 with an R^2 of 0.62; for second-parity cows, the slope was 0.79, and R^2 was 0.71 (Figure 1).

3.3. Evaluation of the Effectiveness of the Correction Application for TDMY Data

The effectiveness of the TDMY data correction was evaluated using 438,973 test records from Validation Set B. It was observed that the average TDMY for each parity significantly increased after correction ($P < 0.05$). For example, for cows with one abnormal record in the first parity, the mean yield increased from 30.25 kg to 30.55 kg ($P < 0.001$); for those with one abnormal record in the second parity, it increased from 34.67 kg to 35.10 kg ($P < 0.001$); and for cows in the third parity and above with one abnormal record, it increased from 35.19 kg to 35.67 kg ($P < 0.001$), as shown in Table 4 Additionally, the coefficient of variation (CV) generally decreased after correction, indicating reduced variability and enhanced stability of TDMY data. This improvement suggested that the correction method effectively minimized anomalous fluctuations, leading to more reliable and consistent measurements.

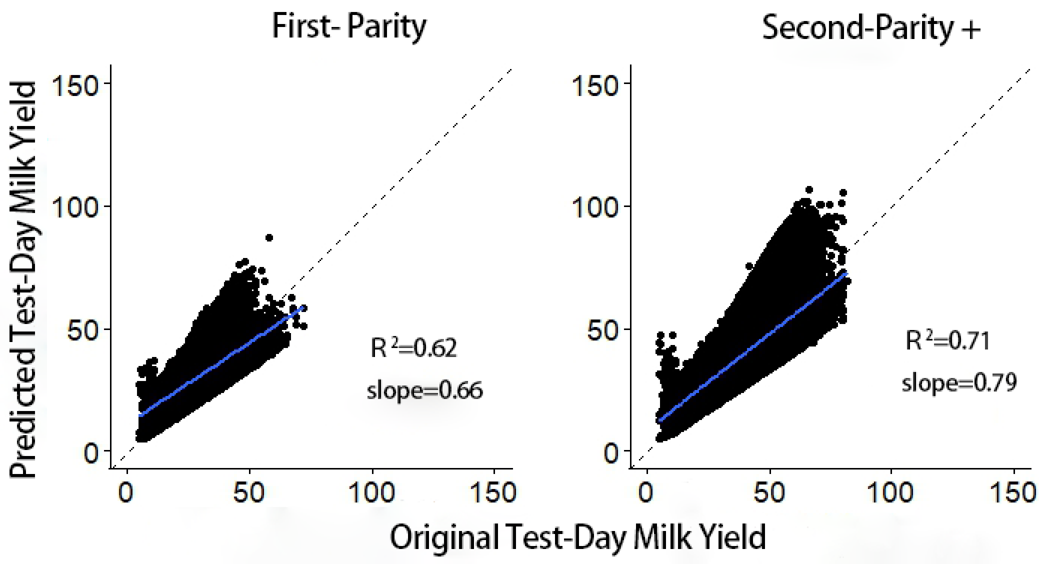

Figure 1. Scatter plot and linear regression of original and predicted TDMY. The Pearson correlation coefficients based on Validation Set A are strong, with values of 0.79 ($p < 0.001$) for first-parity records and 0.84 ($p < 0.001$) for second-parity records.

Table 4. Descriptive statistics of TDMY before and after correction.

Parity	Number of exception records	Sample size	Mean±SD		CV		P-value
			Pre	Post	Pre	Post	
1	1	10,387	30.25±9.84	30.55±9.43	32.54	30.85	< 0.001
	2	3,008	26.80±10.82	26.98±10.00	40.37	37.08	< 0.05
	≥ 3	1,728	24.66±12.29	24.97±10.99	49.85	44.02	< 0.05
2	1	8,263	34.67±12.76	35.10±12.50	36.81	35.61	< 0.001
	2	2,619	30.15±13.12	30.43±12.37	43.53	40.66	< 0.05
	≥ 3	1,626	27.60±14.49	28.01±13.26	52.51	47.34	< 0.01
≥ 3	1	9,894	35.19±13.97	35.67±13.77	39.69	38.60	< 0.001
	2	3,530	30.04±14.28	30.41±13.65	47.53	44.87	< 0.001
	≥ 3	2,242	27.31±15.51	27.74±14.33	56.78	51.67	< 0.01

Note: $P < 0.001$ indicates very high statistical significance, *$P < 0.01$ indicates high statistical significance*, and *$P < 0.05$ indicates statistical significance.*

The 305-day milk yield was calculated using 438,973 test records from Validation Set B to evaluate the effectiveness of TDMY data correction in the calculation of the 305-day yield. The relevant descriptive statistics are shown in Table 5. The 305-day yield calculated after correction increased compared to before correction, with a significant increase ($P < 0.05$) observed in cows with one abnormal record per parity. In contrast, cows with two or more abnormal records showed an increase in their 305-day yield, but this was not significant ($P > 0.05$). The analysis of variance conducted on the 305-day yields before and after correction indicated no significant difference ($P > 0.05$). For cows with two or more abnormal records within the same parity, the coefficient of variation

(CV) for the average 305-day milk yield decreased after data correction. This reduction in CV indicated that the variability of the milk yield data was reduced, reflecting improved data consistency and stability.

Table 5. Mean, standard deviation, and coefficient of variation of 305d milk yield before and after correction at different parities.

	Number of exception records	Sample size	Mean±SD		CV(%)		P-value
			Pre	Post	Pre	Post	
1	1	10,387	9,032.48±2,291.77	9,105.04±2,313.18	25.37	25.41	0.02
	2	3,008	8,011.80±2,383.51	8,054.46±2,331.49	29.75	28.95	0.48
	≥ 3	1,728	7,349.48±2,664.64	7,432.03±2,566.06	36.26	34.53	0.35
2	1	8,263	10,280.73±2,863.20	10,383.88±2,904.36	27.85	27.97	0.02
	2	2,619	8,962.94±2,702.50	9,034.60±2,673.97	30.15	29.60	0.33
	≥ 3	1,626	8,175.82±2,967.29	8,284.19±2,883.42	36.29	34.81	0.29
≥ 3	1	9,894	10,449.50±3,191.36	10,564.47±3,241.70	30.54	30.68	0.01
	2	3,530	8,939.49±3,094.53	9,033.65±3,078.51	34.62	34.08	0.20
	≥ 3	2,242	8,105.70±3,368.72	8,219.94±3,280.85	41.56	39.91	0.25

Note: *$P < 0.05$ indicates statistical significance.*

Based on the 438,973 test-day records from Validation Set B, population-level lactation curves were fitted using the WOOD model, as shown in Figure 2. Table 6 summarizes the estimated parameters (a, b, and c) across parities before and after correction. The results showed that the average TDMY increased for all parities after data correction. Notably, the differences between corrected and uncorrected values were more pronounced before peak lactation, with significantly higher yields observed after correction. Although the differences narrowed after peak lactation, corrected yields remained consistently higher across all stages.

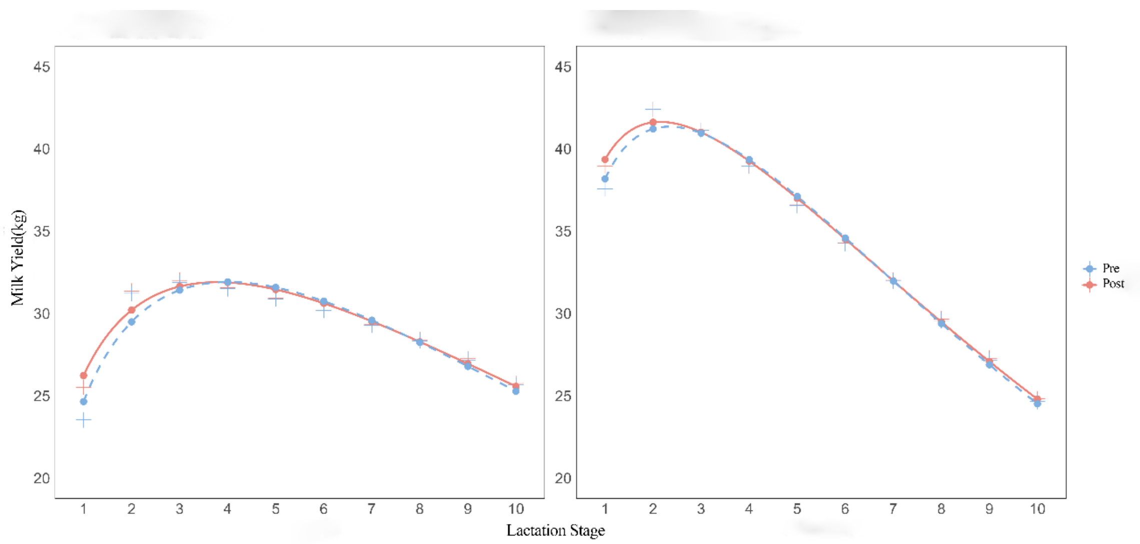

Figure 2. Group lactation curves before and after correction at different parities.

Table 6. Parameters related to group lactation curves before and after correction at different parities.

Item	Lactation 1		Lactation 2+	
	Pre	Post	Pre	Post
a	27.25	28.62	43.16	44.10
b	0.40	0.33	0.29	0.24
c	0.10	0.09	0.12	0.11
R^2	0.92	0.95	0.99	1.00
Per	3.23	3.25	2.70	2.70
PY	31.91	31.90	41.34	41.63
Tmax	4.02	3.78	2.34	2.15

Note: Parameters a, b, and c represent key components of the lactation curve. Based on these parameters, secondary traits such as peak yield, time to peak yield, and lactation persistency were calculated to characterize the lactation performance in more detail.

3.4. Investigation and Analysis of Daily TDMY Records

Statistical analysis was conducted on the proportions of abnormally high and low TDMY records based on different parities and lactation days. In the DHI test records, the proportion of abnormal TDMY records across parities and lactation days ranged from 2.84% to 32.68%, as shown in Figure 3. The proportion of abnormal TDMY records exhibited the following trends: The proportion was higher during early lactation, lower in late lactation but higher than in mid-lactation. The variation trend of abnormal TDMY records in mid-lactation was similar across parities; however, as parity increased, the proportion of abnormal records tended to rise. This increase was particularly significant in records of six parities and above, where the proportion of abnormal TDMY records began to rise significantly during mid-lactation.

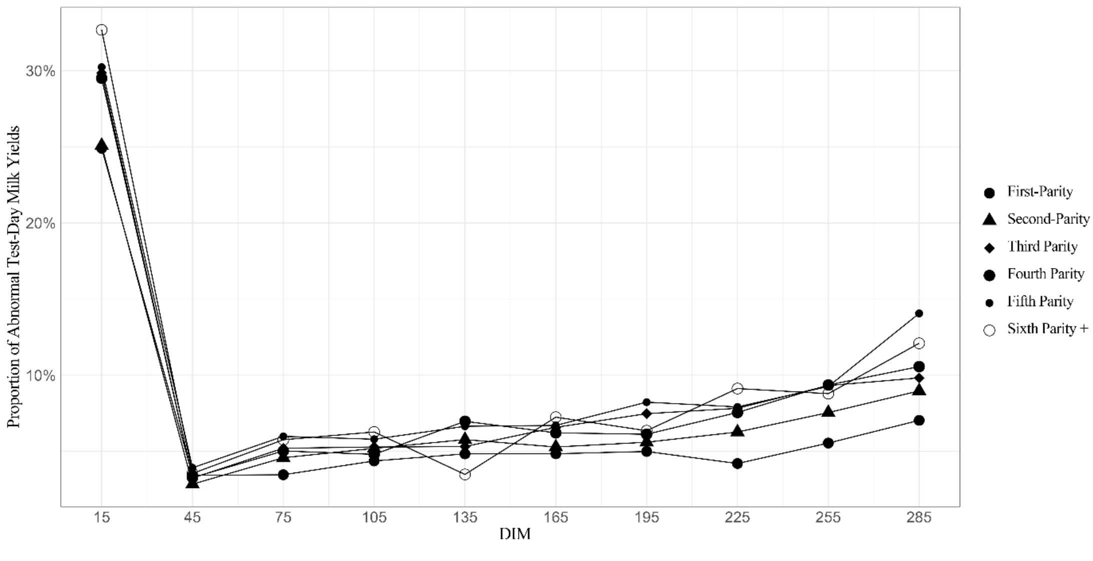

Figure 3. Proportion of test-day records with abnormal TDMY determination across parities and lactation days.

In the DHI test records, only 0.60% to 4.24% of the records were identified as abnormal high TDMY measurements, as shown in Figure 4. The proportion of abnormal high TDMY records for first and second parity cows remained relatively stable throughout the lactation period. In contrast, the proportion of abnormal high records for fifth parity and above exhibited significant variability over the lactation period, particularly showing elevated levels in early and late lactation stages.

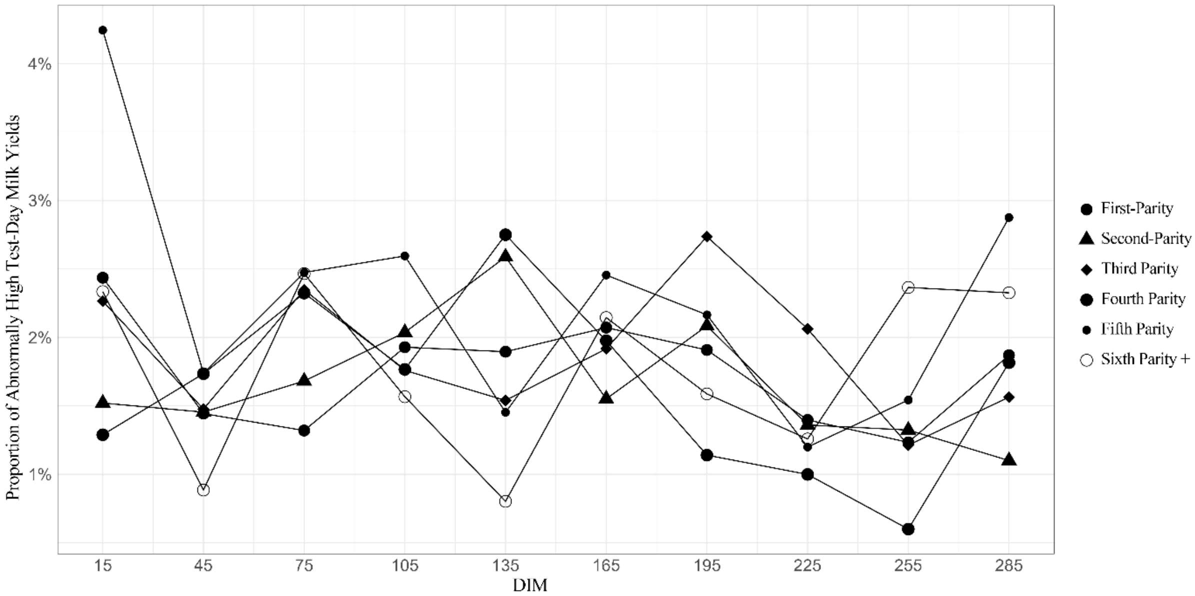

Figure 4. Proportion of test-day records with abnormally high TDMY determination across parities and lactation days.

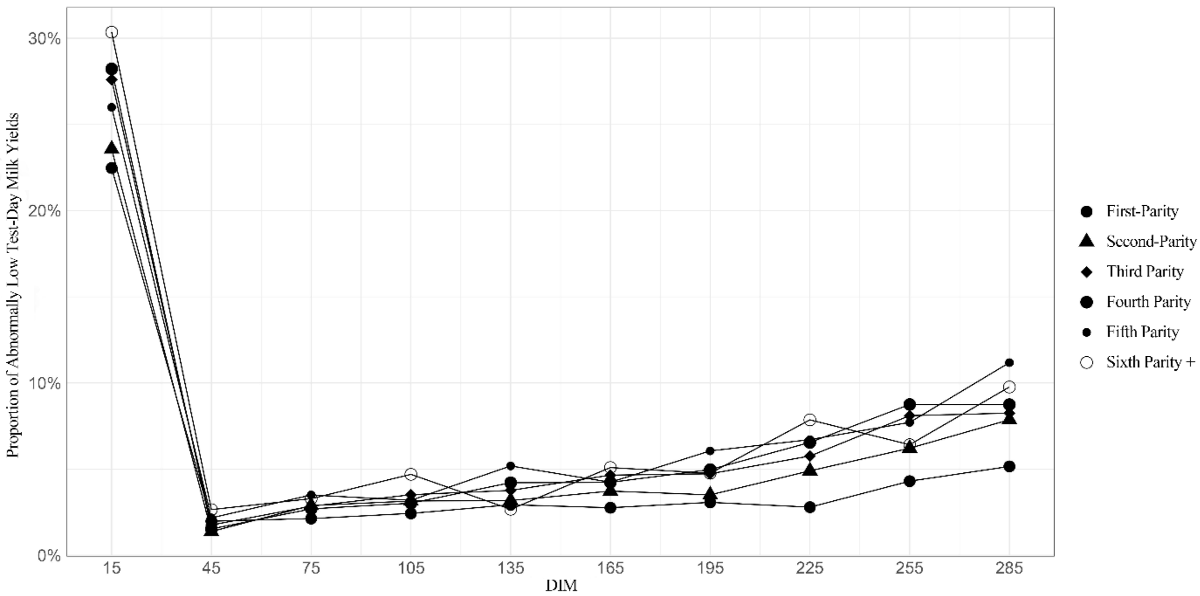

Figure 5. Proportion of test-day records with abnormally low TDMY determination across parities and lactation days.

In the DHI test records, abnormal low TDMY measurements were identified, accounting for 1.38% to 30.35% of the total, as shown in Figure 5. The proportion of these abnormal low measurements was highest in early lactation and gradually decreased as the days in lactation increased. In mid-lactation, the proportion of abnormal low records stabilized, while a slight increase was observed in late lactation. For first parity cows, the proportion of abnormal low TDMY records was relatively low, particularly in mid-lactation, where changes were minimal, with only a slight increase noted in late lactation. Cows from second to fifth parity exhibited a similar and relatively stable trend in the proportion of abnormal low records. However, sixth parity and higher cows showed a higher and more variable proportion of abnormal low TDMY measurements.

4. Discussion

We utilized a substantial amount of TDMY measurement data to estimate the relationship between the slope of TDMY changes, days in lactation, and adjacent TDMY using multiple linear regression. By incorporating information from previous TDMY measurements, adjacent TDMY were predicted. Additionally, different slopes for adjacent TDMY changes were estimated based on various stages of lactation, which helps to more accurately describe the patterns of lactation traits in dairy cows.

In a study conducted in 2002, Wiggans et al. [8] evaluated and corrected records from the U.S. national DHI database, identifying approximately 2% of records as abnormal TDMY measurements. Although the trend of abnormal TDMY measurements across different parities and lactation days was similar to that found in this study, the overall proportion was lower. Gillon et al. [11] proposed a method for corrected best prediction (mBP) to forecast milk yield, while Liseune et al. [12] utilized deep learning techniques to predict milk yield in subsequent lactation periods, demonstrating that predictions made during early lactation were more accurate than those from other forecasting methods.

According to Wiggans et al., the correlation between TDMY measurements before and after adjustment was range 0.93-0.95, accompanied by a slight increase in heritability following the adjustment. In contrast, the present study found that the correlations of adjusted TDMY across parities ranged from 0.79 to 0.84, which is somewhat lower than those reported by Wiggans et al. (2002) [13]. Additionally, Gillon et al. [11] reported correlations ranging from 0.85 to 0.95 between actual and predicted values using corrected best prediction (mBP) for milk yield, which were higher than those observed in this study.

When predicting milk yield using linear regression, outliers significantly impact the results. The extensive time span of the milk yield data used in this study, particularly for earlier records, presented challenges regarding data quality. In particular, the DHI data collected in China over a decade ago suffered from lower data quality in terms of inconsistent or inaccurate on-farm recording of milk yield. Although data quality control procedures were applied when estimating regression coefficients using multivariate models, many data points that did not reflect the true physiological status of dairy cows remained, which may have contributed to reduced prediction accuracy. Cao et al. [14] applied empirical mode decomposition (EMD) to filter out irregular fluctuations from daily lactation milk yield data, which improved data quality and enhanced the accuracy of milk yield predictions using a back propagation neural network. In our study, predictions of TDMY exhibited some degree of error, which could be further reduced by expanding the dataset or incorporating more recent, higher-quality data. Such improvements would enhance the accuracy of prediction and adjustment of TDMY measurements. Here, we focused on the evaluation and correction of TDMY data and its application in lactation curves and 305-day milk yield calculations. We employed the WOOD model to fit individual and population lactation curves, comparing the differences in milk yield before and after correction. The results indicated significant changes in the lactation curves following the corrections. Specifically, the corrected data showed smoother lactation trajectories and more accurate peak yield timing, leading to improved estimation of 305-day milk yield, which better reflects true milk production performance.

After fitting individual lactation curves, the corrected TDMY data retained more records, providing superior data for the study of dairy cow lactation curves. Analysis of the proportion of abnormal TDMY measurements across parities and lactation days revealed that the identification rate of abnormally low TDMY was significantly higher than that of abnormally high measurements throughout the lactation period, particularly in the early and late stages of lactation.

Directly removing these data would lead to information loss, particularly affecting early-lactation or high-yielding cows. Therefore, during the initial quality control process, we identified and corrected these outliers than excluding them, aiming to retain their contribution while minimizing distortion of actual production performance.

This approach enabled us to preserve test-day records that would otherwise have been discarded, reducing bias in the estimation of 305-day milk yield. Consequently, it improved the accuracy of the 305-day yield evaluation and enhanced the stability and reliability of breeding value estimation by increasing data utilization.

5. Conclusions

In this study, we applied multiple linear regression to correct TDMY records from dairy cows in Shandong Province. By adjusting than excluding abnormal values, the approach improved data quality and completeness, particularly in early lactation and high-producing cows. Validation showed reasonable correlations between observed and predicted yields, with better accuracy for later parity cows. The corrections enhanced lactation curve fitting and reduced variation, contributing to more accurate 305-day milk yield estimates. These improvements demonstrate practical accuracy that supports more reliable genetic evaluations and improved herd management.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the Earmarked Fund for the China Agriculture Research System (CARS-36), the National Agriculture Key Science & Technology Project (NK20221201), Agricultural Scientific and Technological Innovation Project of Shandong Academy of Agricultural Sciences (CXGC2025F09), and the National Key R&D Program of China (2021YFF1000701-06). The authors thank Shandong OX Livestock Breeding Co., Ltd., Shandong Province, China, for providing DHI data used in this study.

Conflict of interest

The authors declare that they have no conflicts of interest. Jianbin Li is an editorial board member for *AIMS Animal Science* and was not involved in the editorial review or the decision to publish this article.

References

1. Hudson CD, Green MJ (2018) Associations between routinely collected Dairy Herd Improvement data and insemination outcome in UK dairy herds. *J Dairy Sci* 101: 11262–11274. https://doi.org/10.3168/jds.2017-13962
2. Koorn D (1998) The introduction of a milk recording system based on milk yields of ICAR approved electronic milk meters and AT4 sampling for fat, protein and SCC. *Public-Europ Assoc Anim Product* 91: 147–150. https://doi.org/10.1163/9789004684010_026
3. Bünger A, Kuwan K, Reinhardt F, Brahmstaedt H-U, Reents R (2005) *Performance Recording of Animals-State of the Art, 2004*. Wageningen Academic, 229–235. https://doi.org/10.3920/9789086865369_038
4. Gao H, Madsen P, Pösö J, Aamand G, Lidauer M, Jensen J (2018) Multivariate outlier detection for routine Nordic dairy cattle genetic evaluation in the Nordic Holstein and Red population. *J. Dairy Sci.* **101**: 11159–11164. https://doi.org/10.3168/jds.2018-15123
5. Madsen P, Pösö J, Pedersen J, Lidauer M, Jensen J (2012) Screening for outliers in multiple trait genetic evaluation. *Interbull Bulletin* 46.
6. Lamanna M, Bovo M, Cavallini D (2025) Wearable collar technologies for dairy cows: A systematized review of the current applications and future innovations in precision livestock farming. *Animals* 15: 458. https://doi.org/10.3390/ani15030458
7. Cavallini D, Giammarco M, Buonaiuto G, Vignola G, De Matos Vettori J, Lamanna M, et al., (2025) Two years of precision livestock management: Harnessing ear tag device behavioral data for pregnancy detection in free-range dairy cattle on silage/hay-mix ration. *Front. Animal Sci.* 6: 1547395. https://doi.org/10.3389/fanim.2025.1547395
8. Wiggans G, Vanraden P, Philpot J (2003) Detection and adjustment of abnormal test-day yields. *J Dairy Sci* 86: 2721–2724. https://doi.org/10.3389/fanim.2025.1547395
9. Group IDCMRW (2017) Section 2-Guidelines for Dairy Cattle Milk Recording.
10. Wood P (1967) Algebraic model of the lactation curve in cattle. *Nature* 216: 164–165. https://doi.org/10.1038/216164a0
11. Gillon A, Abras S, Mayeres P, Bertozzi C, Gengler N (2010) Adding value to test-day data by using modified best prediction method. *ICAR Technical Series* 14.
12. Liseune A, Salamone M, Van Den Poel D, Van Ranst B, Hostens M (2021) Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning. *Comput Electron Agri* 180:105904. https://doi.org/10.1016/j.compag.2020.105904
13. Wiggans G, Vanraden P, Bormann J, Philpot J, Druet T, Gengler N (2002) Deriving lactation yields from test-day yields adjusted for lactation stage, age, pregnancy, and herd test date. *J Dairy Sci* 85: e261–264. https://doi.org/10.3168/jds.S0022-0302(02)74077-0
14. Cao Z, Zhao H, Xu J, Zhang G, Li Y, Su Y, et al. (2022) Using empirical modal decomposition to improve the daily milk yield prediction of cows. *Wireless Commun Mobile Comput* 1: 1685841.

https://doi.org/10.1155/2022/1685841