Contents lists available at ScienceDirect

Animal

The international journal of animal biosciences

Prediction of first test day milk yield using historical records in dairy cows

M. Salamone[a,b,*], I. Adriaens[b], A. Vervaet[a], G. Opsomer[a], H. Atashi[c], V. Fievez[d], B. Aernouts[b,1], M. Hostens[d,e,1]

[a] Department Internal Medicine, Reproduction and Population Medicine, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium
[b] Department of Biosystems, Division of Animal and Human Health Engineering, KU Leuven, Campus Geel, Kleinhoefstraat 4, 2440 Geel, Belgium
[c] Department of Animal Science, Shiraz University, Shiraz, Iran
[d] Department of Animal Sciences and Aquatic Ecology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Ghent, Belgium
[e] Department of Population Health Sciences, Division of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 7, 3584 CL Utrecht, the Netherlands

ARTICLE INFO

Article history:
Received 3 September 2021
Revised 15 September 2022
Accepted 16 September 2022
Available online 17 October 2022

Keywords:
Dairy cow
Health monitoring
Milk yield prediction
Random forest model
Transition period

ABSTRACT

The transition between two lactations remains one of the most critical periods during the productive life of dairy cows. In this study, we aimed to develop a model that predicts the milk yield of dairy cows from test day milk yield data collected in the previous lactation. In the past, data routinely collected in the context of herd improvement programmes on dairy farms have been used to provide insights in the health status of animals or for genetic evaluations. Typically, only data from the current lactation is used, comparing expected (i.e., unperturbed) with realised milk yields. This approach cannot be used to monitor the transition period due to the lack of unperturbed milk yields at the start of a lactation. For multiparous cows, an opportunity lies in the use of data from the previous lactation to predict the expected production of the next one. We developed a methodology to predict the first test day milk yield after calving using information from the previous lactation. To this end, three random forest models (nextMILK$_{FULL}$, nextMILK$_{PH}$, and nextMILK$_{P}$) were trained with three different feature sets to forecast the milk yield on the first test day of the next lactation. To evaluate the added value of using a machine-learning approach against simple models based on contemporary animals or production in the previous lactation, we compared the nextMILK models with four benchmark models. The nextMILK models had an RMSE ranging from 6.08 to 6.24 kg of milk. In conclusion, the nextMILK models had a better prediction performance compared to the benchmark models. Application-wise, the proposed methodology could be part of a monitoring tool tailored towards the transition period. Future research should focus on validation of the developed methodology within such tool.

Implications

In the present study, we explored the potential of historical milk test day data to predict the individual daily milk yield on the first test day of the next lactation. The results show that the utilisation of such historical data seems to allow accurate prediction of production at the start of the next lactation when compared with benchmark models. These models could ultimately be used in a wide range of applications, from economic evaluation by expanding the forecast horizon of farmers to the next lactation, to the implementation as data-driven health monitoring tools by comparing the expected production with the realised production.

Introduction

The increase in milk yield of dairy cattle coincides with multiple challenges imposed on the cows, especially during the transition period in the six weeks around calving Journal of Dairy Science, 101(10), 9419–9429 (Probo et al., 2018). In literature, the duration of transition period has been argued in recent years, certain authors define the transition period from dry off till 6 weeks after calving (Lopreiato et al., 2020). In this period, 30–50% of the cows develop metabolic or infectious diseases such as mastitis, metritis, ketosis, lameness or displaced abomasum (LeBlanc, 2010; Hostens et al., 2012; Pascottini et al., 2020). To better support animals at

[*] Corresponding author at: Department Internal Medicine, Reproduction and Population Medicine, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.

E-mail address: Matthieu.Salamone@ugent.be (M. Salamone).

[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.animal.2022.100658
1751-7311/© 2022 The Authors. Published by Elsevier B.V. on behalf of The Animal Consortium.

risk of developing transition problems, identification methods to date have mainly focused on the laboratory analysis of metabolic markers as discriminating factor, either requiring milk or blood samples (Saun and Robert, 2006; De Koster et al., 2019; Grelet et al., 2019). However, these techniques require action from the farmer, which may lead to poor identification performance. The development of an automatic alert system to point out animals at risk could drastically improve the early identification of sick animals.

The advent of routine data collection on dairy farms has led to the opportunity to develop data-driven and automated health monitoring tools. For example, Adriaens et al. (2021) proposed to monitor udder health through perturbations in milk yield. A bottleneck to this approach is that it requires an accurate estimation of the expected milk yield in an unperturbed state, typically derived from a theoretical lactation shape and the production data available from a certain lactation (Poppe et al., 2020; Adriaens et al., 2021; Ben Abdelkrim et al., 2021). In this context, multiple data-based models using high-frequency milk meter data have been developed to predict the unperturbed milk yield within the same lactation (Macciotta et al., 2011; Adriaens et al., 2018). Up till today, these models have had a wide range of applications. They have, for instance, been used to estimate the expected production at herd or cow level, as tools to design suitable breeding strategies or genetic selection criteria, as individual health monitoring algorithms and as tools that estimate the response to management and environmental changes (Dematawewa et al., 2007; Ehrlich, 2011; Macciotta et al., 2011). Predicting the unperturbed milk production in the first weeks of lactation, however, is challenging, because at that stage, there is too little milk production data to accuratly fit a lactation curve model on. More specifically, estimates of the unperturbed state based on only the first few days of milk production are deemed unreliable because health problems might already have influenced the lactation performance at or even before the start of the lactation. Therefore, differences between predicted milk yield and actual milk yield could be low while health issues remain unnoticed in the transition period. Nordlund (2006) identified a potential solution which could overcome the aforementioned issues, by using information from the previous lactation to make predictions on the milk production of the next lactation.

The application of advanced machine-learning techniques to the increasing amount of data in dairy farming has led to the development of new insights into animal welfare and to real-time monitoring possibilities (Hermans et al., 2018). For example, recently, Liseune et al. (2021) presented a deep-learning model to predict the unperturbed daily milk yield of the first 305 days in milk (**DIM**), using daily milk meter data of the previous lactation in combination with cow and herd key performance indicators. This model demonstrated the application of machine learning to predict milk production of the next lactation. Relying on daily sensor data, this model excludes a wide range of farms that do not have milk meters installed. An alternative approach is to use test day records (**TDRs**) collected through national dairy herd improvement programmes. TDRs are recorded with a frequency of four to eight weeks, but present the advantage of having been collected for many years on a majority of dairy farms, thus having plenty of historical data readily available.

In the past, models based on TDR have been proposed for the genetic evaluation of dairy cattle as a replacement for the traditional 305-d lactation yield. These models typically have the ability to account for environmental effects occurring on the day of milk recording (Mayeres et al., 2004). Additionally, random regression test day models have been developed to predict the future performance of animals in the same parity (for example, milk yield in the ninth month of the lactation based on the performance of animals in the first eight months of the current lactation). To our current knowledge, no study has evaluated the potential of using TDR of the current lactation to predict the performance of an animal in a next lactation. As TDRs are widely available and standardised, they allow the training of complex machine-learning models. In their turn, these models can create added value to data that is already routinely collected on dairy farms.

In the present study, we aimed to combine powerful machine learning techniques with the idea of using historical data to predict milk production in the next lactation. To this end, a set of random forest regression models was developed to predict the expected milk yield on the first test day of a lactation, based on features derived from historical TDR of the previous lactation and additional cow and herd information.

Material and methods

Raw data

The raw dataset was accessed via the MmmooOgle™ platform (Bovicom, Puurs, Belgium) and originated from 102 herds located in six countries (BE: n = 74, the NL: n = 16 DE: n = 5, USA: n = 4, FR: n = 2, IT: n = 1), spanning a period of 20 years between 2000 and 2020. In total, data from 83 406 animals with on average 2.6 lactations per animal were available. These data included TDR and the corresponding cow information such as cow identifiers, calving dates, breeding dates and dry-off dates. At herd level, the recording of data started between 2000 and 2012 and ended between 2007 and 2020, with a median time span of 18.1 years (Q1: 15.8 years, Q3: 20.1 years). The TDRs contained the following information: daily milk yield, test date, parity and DIM. Additionally, the data contained lactation curve parameters of the MilkBot model fitted on the test day milk yield (Ehrlich, 2011). These MilkBot parameters (scale, offset, ramp and decay), summarise the shape of the lactation curve in a standardised way. A general flowchart displaying the different steps presented in this methodology can be found in Fig. 1.

Data selection

All data editing, data processing and the model development were done using Apache Spark version 3.0.0 (The Apache Software Foundation, Wakefield, USA) running on the high-performance computing infrastructure of Ghent University, Belgium.

In order to obtain data suitable for the analysis, e.g. in terms of completeness, several selection steps were implemented. First, lactations with missing MilkBot parameters were removed. These parameters were essential to calculate cumulative milk yield at animal and herd levels. Second, lactations followed by a subsequent lactation with at least one TDR were selected, referred to as lactation *X* and lactation *X + 1,* respectively. The milk yield (in kg) in that first TDR of lactation *X + 1* ($\mathbf{kgTD1}_{X+1}$) was defined as the dependent variable to be predicted.

The second selection step aimed at identifying lactations for which data quality was sufficient for the analysis. Moreover, lactations *X + 1* where the first test day ($\mathbf{TD1}_{X+1}$) was measured after 60 DIM were excluded. Accordingly, the range for the first test day of the next lactation for which the model was trained and validated encompasses the period in which high-producing dairy cows face major transition challenges (LeBlanc, 2010; Probo et al., 2018; Lopreiato et al., 2020).

Furthermore, the calving interval in lactation *X* was at least 300 days and at most 530 days. The minimal and maximal age at first calving had to be between respectively 20 months and 36 months. Additionally, lactations with less than eight TDR in lactation *X* were excluded to ensure having sufficient data to make

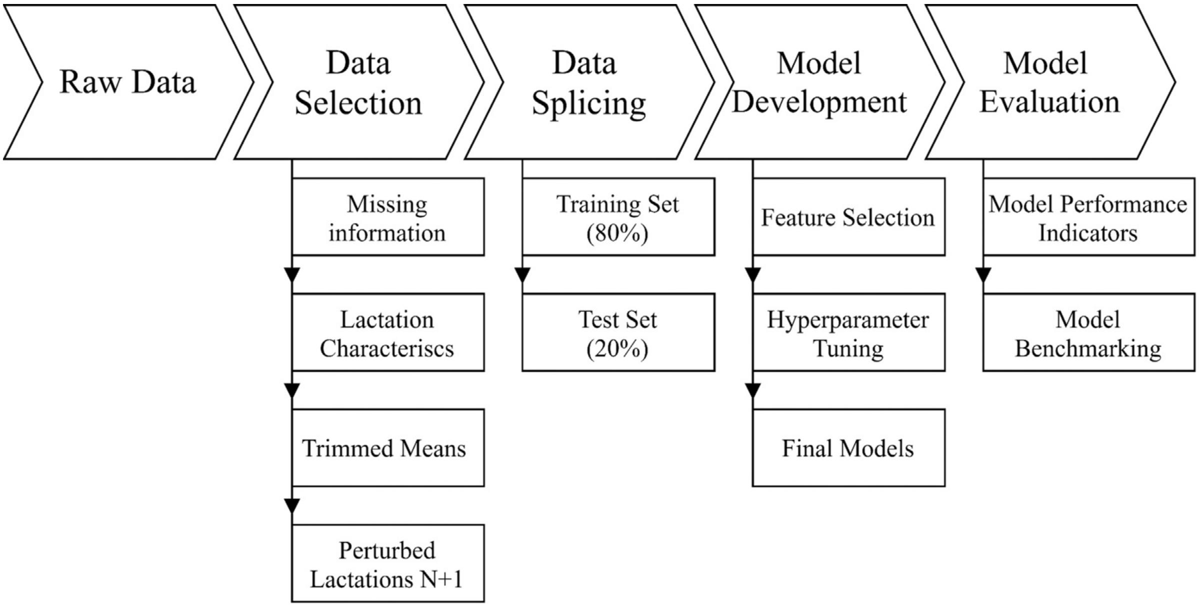

Fig. 1. Flowchart displaying the different steps applied in this study to develop the nextMILK models for dairy cows.

the predictions. When more than eight TDRs were available, the TDRs after the 8th were excluded from the predictors. The eight selected TDRs had to be taken regularly throughout the lactation, for example every 4–8 weeks. Lactations *X* for which no regularity between TDR was found were removed from the analysis.

The third and final selection step was applied at the lactation level and aimed at identifying lactations from which the first test day yield potentially belonged to a milk yield perturbation using characteristics of the dependent variable kgTD1$_{X+1}$. Because the aim of the model is to predict the expected milk yield of healthy cows in the beginning of lactation *X* + 1, all 'unhealthy' lactations *X* + 1 were removed from the dataset. Because detailed health information of the cows is not available for this dataset, a lactation *X* + 1 is 'unhealthy' if it is perturbed on the first test day relative to the MilkBot lactation model fitted to the milk yield data of the entire lactation *X* + 1. Moreover, if the actual daily milk production on the first test day of lactation *X* + 1 was more than 6 kg below the expected milk yield according to the MilkBot model, then this lactation was identified as being 'unhealthy' and thus rejected. The 6 kg threshold corresponds with the reported RMSE performance of the MilkBot model in second and greater parities (Cole et al., 2012). In Fig. 2, a flowchart shows the effect of all these data selection steps on the number of lactations and animals present in the final dataset.

Feature description

Several features at different levels (test day, lactation, cow or herd) were available or could be calculated from the available data to enter into the prediction models. In this section, an overview of these features is given. These features included the DIM and corresponding milk yields at each test day of lactation *X* and the DIM of TD1$_{X+1}$. Features at the lactation level were parity number, the lactation length, days open for lactation *X*, the calving interval and days dry between lactation *X* and *X + 1*. Parity number was considered as an ordinal categorical variable; parity number 5 and above were grouped in the same category. The age at first calving was considered as a cow-specific feature.

A set of additional features were derived from the features presented above. At the lactation level, six seasons of calving classes were defined by the month of calving to take into account the seasonality of a lactation. The first class was attributed to December and January; the next classes were attributed to each consecutive pair of months. The season of calving was defined for both lactation *X* and lactation *X + 1*. The cumulative milk yield at 21 DIM (**M21**), 75 DIM (**M75**) and 305 DIM (**M305**) of each lactation *X* was calculated by summing the daily milk yields for the different time spans using the MilkBot equation and parameters.

Some additional features describing the herd performance were calculated, including the average cumulative milk yields (M21, M75, M305), average age at first calving, average calving interval, days dry, lactation length, days open and the maximal and minimal daily milk production during the lactation. To account for different production levels in different lactations, these herd averages were stratified for each parity. Furthermore, to account for evolutions in production in a herd over time, the averages were also calculated in function of time using a sliding window of two years. The resulting herd average features were used to compute absolute differences of a cow in comparison with the herd accounting for the respective herd, year and parity number. Year was defined as the year in which lactation *X* started.

Data splicing

The aforementioned final dataset was randomly spliced into a test set (20%) and a training set (80%). Data were split before any model development was done to create an independent test dataset to evaluate the final model. The data splicing was performed randomly at animal level, preventing the inclusion data of the same cow in both test set and training set. The usage of absolute differences with regard to the herd performances described in the previous section preserves the independence between both sets.

Model development

In the present study, we developed a random forest regression model to predict the first test day milk yield in lactation *X + 1* from

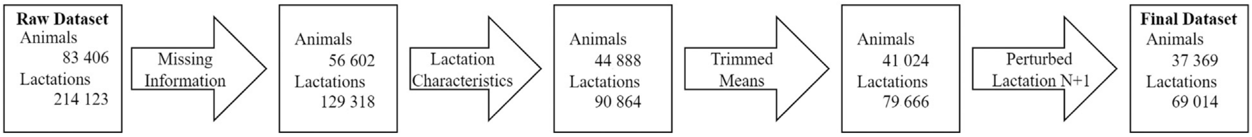

Fig. 2. Graphical representation of the effects of each filtration step on the number of lactation and the number of distinct cows. This in order to construct the final dataset.

the above-mentioned features. The following paragraphs describe the steps taken to train and evaluate the model.

Model description

The random forest regression models were trained on the training set using the native MLlib library function of Apache Spark (The Apache Software Foundation, Wilmington DE, USA). Random forest regression is a supervised ensemble learning method proposed by Breiman (2001). These models consist of multiple decision trees, where the prediction of each constituting decision tree is combined to create a final prediction. Random forest models can be used in classification or regression applications. When the random forest is used for regression, the predictions of the individual decision trees are averaged to obtain the final estimation. Individual decision trees have an inherent tendency to overfit training data, random forest models mitigate this by combining the prediction of a multitude of individual decision trees reducing the overfitting problem (James et al., 2013). Additionally, when compared to black box models such as artificial neural networks, where gathering insights on the model's functioning is rather difficult, random forest provides the possibility to extract feature importance metrics. These metrics represent the relative importance of a feature normalised to sum to 1 calculated by the method presented in Hastie et al. (2009). In dairy research, random forest models have been mainly used as a classifier (Walsh et al., 2007; Shahinfar et al., 2014; Parker Gaddis et al., 2016; Borchers et al., 2017; De Koster et al., 2019). However, in recent years, the usage of random forest models in regression applications for livestock data has increased (Dallago et al., 2019; van der Heide et al., 2019). The development of the random forest regression model consisted of the following steps: feature selection, defining the optimal hyperparameters, training of the model to obtain model parameters and evaluation of the final model.

Feature set selection

The full feature set comprised all 40 features described above. We defined three feature categories within this final feature set: (1) the individual production features, (2) herd-level production features and (3) reproduction-derived features (age at first calving, calving interval, days dry, days open, lactation length and the absolute difference of those features with the herd average). From this point on, three sets of features were used: only the production features (**P**), production and herd features (**PH**) and finally the full feature set (**FULL**). A graphical representation of those three sets and their composing features is provided in Fig. 3.

Hyperparameter tuning

In this step, for each feature set, the optimal pair of hyperparameters was established, which consists of the number of composing trees (5, 25, 125, 250 and 500 trees) and the maximal depth of those trees (5, 10, 15, 20, 25 and 30 levels deep). A subset of the random forest models was trained by using a random 5-fold cross-validation on the training set.

The model performance was evaluated using the error with which the model is able to predict kgTD1$_{X+1}$. For this, RMSE is best suited, with lower RMSE being better. This performance metric is calculated by taking the square root of the average squared difference between the actual kgTD1$_{X+1}$ (y_i) and the predicted kgTD1$_{X+1}$ ($\hat{y}_i$). The RMSE is expressed as seen in Eq. (1), with N the total number of predicted values:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{N}} \tag{1}$$

The RMSE represents an absolute error with a clear unit (kg of milk) and the interpretation is straightforward: RSME is a positive value, where the closer to 0, the better the model is able to predict the dependent variable. During this step, the best hyperparameter set model was defined by the lowest RMSE in the cross-validation. Statistically significant differences between the best-performing hyperparameter set and the other sets were calculated on the squared value of the residuals by applying one-sided paired *t*-tests using SparkR. If multiple models had squared residuals that were not statistically higher than the best-performing model, the least complex structure of these models was chosen.

Final models

This optimal set of hyperparameters was then used to train the final random forest models, referred further as nextMILK$_{\mathrm{FULL}}$, nextMILK$_{\mathrm{PH}}$, nextMILK$_{\mathrm{P}}$ models using the respective feature set. They were trained with the full training set (80% of the lactation pairs) and evaluated by predicting the kgTD1$_{X+1}$ for the test set. The feature importance in function of the predicted kgTD1$_{X+1}$ was also analysed, to identify any biases or inconsistencies in the model performances. Ultimately, a set of model performance indicators were computed, to complete the evaluation of the nextMILK models.

Model evaluation

Model performance indicators

The performance of the final nextMILK models was evaluated on the test set using four Model Performance Indicators (**MPIs**) commonly used in similar studies, including the RMSE. In addition, mean absolute error (**MAE**), mean percentage error (**MAPE**) and R^2 were also used to evaluate the final model.

The MAE represents the average absolute difference between the actual kgTD1$_{X+1}$ (y_i) and the predicted kgTD1$_{X+1}$ ($\hat{y}_i$), represented in Eq. (2):

$$MAE = \frac{1}{N}\sum_{i=1}^{N-1}\left|y_i - \hat{y}_i\right| \tag{2}$$

Even though the definition and interpretation of MAE are similar to RMSE, MAE is less influenced by outliers in the residuals. More concretely, in addition to identifying the model with lower errors overall, RMSE provides a better view on which model is less sensitive to extreme values in the prediction errors. For this reason, MAE will always be lower than RMSE, which could lead to overestimations of model performance in the case of large variation in the residuals.

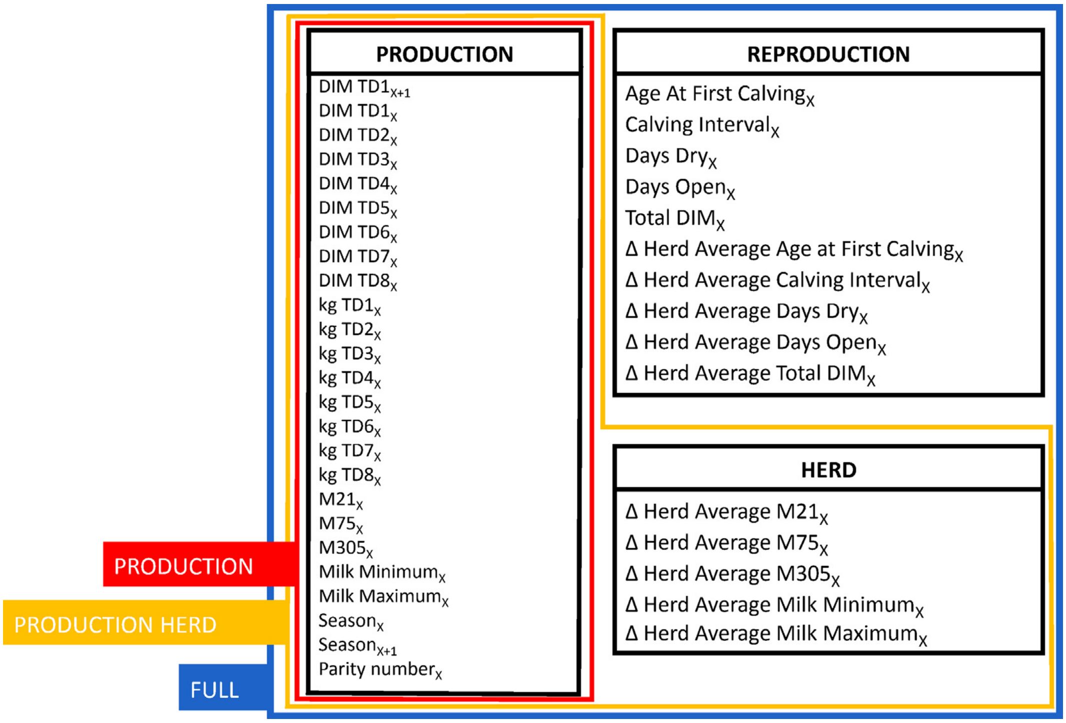

Fig. 3. Graphical representation of the cow features within their classes and the three feature sets created with each of those classes. Abbreviations: DIM = days in milk, TD = test day, M = cumulative milk yield, *X* = lactation *X*, *X* + 1 = lactation *X* + 1.

The MAPE is calculated as the difference between actual kgTD1$_{X+1}$ (y_i) and the predicted kgTD1$_{X+1}$ ($\hat{y}_i$), divided by the actual kgTD1$_{X+1}$ as shown in Eq. (3):

$$MAPE = \frac{1}{N}\sum\nolimits_{i=1}^{N-1}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3}$$

This MPI is a relative value between 0 and 1, where lower values indicate better predictions. It has the advantage of displaying errors in function of the actual value. However, MAPE penalises negative errors more than positive errors, causing it to be a metric biased to favour models which underestimate the dependent variable.

The R^2 is calculated as shown in Eq. (4). The upper term represents the residual sums of squares, where difference between the actual kgTD1$_{X+1}$ (y_i) and the predicted kgTD1$_{X+1}$ ($\hat{y}_i$) is squared. The lower term represents the total sum of squares, which is the difference between the actual kgTD1$_{X+1}$ (y_i) and the overall mean of the actual kgTD1$_{X+1}$ ($\bar{y}_i$):

$$R^2 = 1 - \frac{\sum_{i=1}^{N-1}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N-1}\left(y_i - \bar{y}_i\right)^2} \tag{4}$$

Model benchmarking

Besides the evaluation of prediction performance, benchmarking a newly developed ML model (i.e., the ones developed by us) against simpler methods can help to assess the added value of this work. To this end, we defined expert-based benchmark models that predict the kgTD1$_{X+1}$ from (1) herd averages, and (2) the lactation curve of lactation X. The herd average milk yield for a certain test day was calculated at year-parity level, using information from the two previous years. For example, if lactation X started in 2012, the data from 2010 to 2011 of that herd were used. More specifically, two types of benchmark models were defined, the HERD-HERD and ANIMAL-HERD models.

- For the "HERD-HERD" benchmark models, the kgTD1$_{X+1}$ was predicted from the herd milk yield, taking the difference in milk yield between parities during (1) the first 75 days of lactation; HERD-HERD-75 or (2) during 305 days of lactation; HERD-HERD-305 into account. For example, if the relative milk yield for parity 2 was on average 20% higher compared to parity 1 for a herd-year and the MilkBot model estimated that a cow produced 20 kg of milk on DIM of TD$_X$ during the first lactation, the kgTD$_{X+1}$ for that lactation was predicted as 20 kg × 120% = 24 kg.
- For the ANIMAL-HERD models, also the performance of the animal compared with the contemporary herd mates was included. More specifically, when the animal produced e.g. 30% less than the contemporary herd mates in lactation X on TD$_X$, this factor was taken into account for the ANIMAL-HERD benchmark models. For example, a cow producing 20 kg of milk in lactation *X* for which the herd produced on average 22 kg on that same test day, and for which parity 2 of the herd produced on average 25% more compared to the first parity, was predicted to produce 20/22 * 1.25 * 20 kg = 22.7 kg. For the herd effect, also here two distinct models were defined, taking either the first 75 DIM of lactation (ANIMAL-HERD–75) and 305 DIM (ANIMAL-HERD–305) into account.

The benchmark models and the nextMILK model were used to predict the kgTD1 $_{X+1}$ on the same test set. From these models, the residuals between predictions and actual kgTD1$_{X+1}$ were calculated to evaluate the model performance using the MPI. To compare the nextMILK and benchmark models statistically, the squared residuals were compared using ANOVA followed by a Benjamini & Yekutieli (2001) corrected paired one-sided *t*-test, significance was defined by *P* < .05.

Results

Data descriptive

After the data selection steps, the dataset comprised 102 distinct herds with in total of 37 369 unique cows and 69 014 lactation pairs. A general descriptive summary of the dataset after selection can be found in Table 1, together with an overview of general production and reproduction characteristics. The average kg of milk produced at TD1$_{X+1}$ was equal to 38.4 ± 8.6 (mean ± SD). The median DIM at which the TD1$_{X+1}$ took place was 21 (Q1: 13 days, Q3: 29 days). Fig. 4a shows the distribution of DIM at TD1$_{X+1}$, demonstrating that the number of $X + 1$ lactations with TD1$_{x+1}$ after 30 DIM is drastically reduced. In Fig. 4b, the average milk production in function of the DIM at TD1 is plotted, from which the low typical production at the start of the lactation can be seen. The split of the final data set yielded a training set and test set counting respectively 57 282 lactations and 11 732 lactations.

Hyperparameter tuning

A total of 450 (five composing trees * six maximal depth of trees * five cross-validation steps * three feature sets) subset models were trained for hyperparameter tuning. The performances of each of those models are displayed in Fig. 5 for each of the three feature sets. When analysing the performances for the P feature set, the hyperparameter combination which yielded the best performance was 500 trees with a maximal depth of 30. The one-sided *t*-tests pointed out the optimal feature set as the combination of 125 trees with a maximal depth of 20. This combination showed no significant difference with the best-performing hyperparameter set. Similar results were found for the PH feature set and the FULL feature set, where in both cases, the best-performing model was constructed by 500 trees with a maximal depth of 30. The optimal combination was found by the *t*-tests to be 125 trees and a maximal depth of 15. These optimal hyperparameters were used in the rest of this study.

Model performances

The MPI values of the final models computed for the complete test set are summarised in Table 2. The full feature set yielded the lowest RMSE, MAE and MAPE and the highest R^2, though the residuals were not significantly different. In Fig. 6, the phenomenon of regression to the mean represented by all the nextMILK models is shown. The SD of the dependent variable kgTD1$_{X+1}$ in the test set is equal to 8.79 kg while for the models, a SD of 5.97, 5.88 and 6.04 kg was found for respectively nextMILK$_{\text{FULL}}$, nextMILK$_{\text{PH}}$, nextMILK$_{\text{P}}$. In Fig. 7, the performance of the nextMILK models is plotted in function of DIM TD1$_{X+1}$. It displays an aspect of the models' performances, where fewer observations of lactations within the range 0–5 DIM TD1$_{X+1}$ and 50–60 TD1$_{X+1}$ presented in Fig. 4a seem to result in a higher variation in RMSE.

The feature importance has been extracted for each of the final models; the top 10 most important features are displayed in Table 3. The consistent presence of the same top five features in all the feature importance lists combined with their high importance score emphasises the importance of these features. The most important feature is DIM TD1$_{X+1}$ in all the three feature sets, followed by the M305 of the lactation *X*. The milk production at the 4th and 5th TD is also found to be consistently present in the top five of most important features. It can be noted that for those TD$_X$ production features, the corresponding DIM$_X$ represents the tail of the feature importance list in all three final models. Furthermore, the herd and reproduction parameters generally have a relatively low importance in the respective nextMILK models.

Model benchmarking

We identified that each of the benchmark models had significantly higher (P < .05) residuals (and thus, a lower prediction performance) compared to the three nextMILK models. Additionally, no significant difference in residuals between the nextMILK models was found. These results show the added value of our method compared to less complicated benchmark models. These results are summarised in Table 2 presenting also the MPI of both benchmark and nextMILK models. The difference in RMSE between the nextMILK models and the benchmark models ranges from 1.65 to 1.26 kg; in percentage, this difference ranges from 23 to 15%. This difference is not present in MAE where nextMILK$_{\text{P}}$ and nextMILK$_{\text{PH}}$ had similar MAE to benchmark HERD-HERD – 75, ANIMAL-HERD – 75 and ANIMAL-HERD – 305.

Discussion

Overall, the calculated performance of the nextMILK models demonstrates the potential of using historical production data to predict milk production in the early stage of the next lactation. The consistent importance of DIM TD1$_{X+1}$ seem to indicate that

Table 1
Overview of available dairy cow data in the final dataset after the data selection step.

Item	Number for the whole dataset	Mean ± SD over herds	Range over herds [minimum; maximum]
Number of herds	102		
Number of cows	37 369	366 ± 365	[3; 2 254]
Number of lactation	69 014	677 ± 654	[4; 4 361]
Parity 1	26 097	256 ± 285.0	[2; 1 726]
Parity 2	19 280	189 ± 189	[2; 1 285]
Parity 3	11 778	117 ± 104	[1; 649]
Parity 4	6 268	65 ± 54	[1; 327]
Parity 5+	5 559	56 ± 55	[1; 285]
Age at first calving (years)		2.1 ± 0.2	[1.7; 3.0]
Interval between TD (days)		33.5 ± 5.1	[23; 56]
Calving interval (days)		397.5 ± 49.1	[300; 529]
305d milk yield (kg)			
Parity 1		8 489 ± 1 732	[3 034; 17 916]
Parity 2		9 835.0 ± 2 034	[2 676; 22 030]
Parity 3		10 304 ± 2 039	[2 628; 20 920]
Parity 4		10 294 ± 2 034	[3 179; 20 326]
Parity 5+		9 971 ± 1 923	[3 194 22 403]
Dependent variable: kgTD1$_{X+1}$ (kg)		38.4 ± 8.6	[3.74; 75.32]

Abbreviations: TD = test day, kgTD1$_{X+1}$ = milk yield of the first test day in lactation $X + 1$.

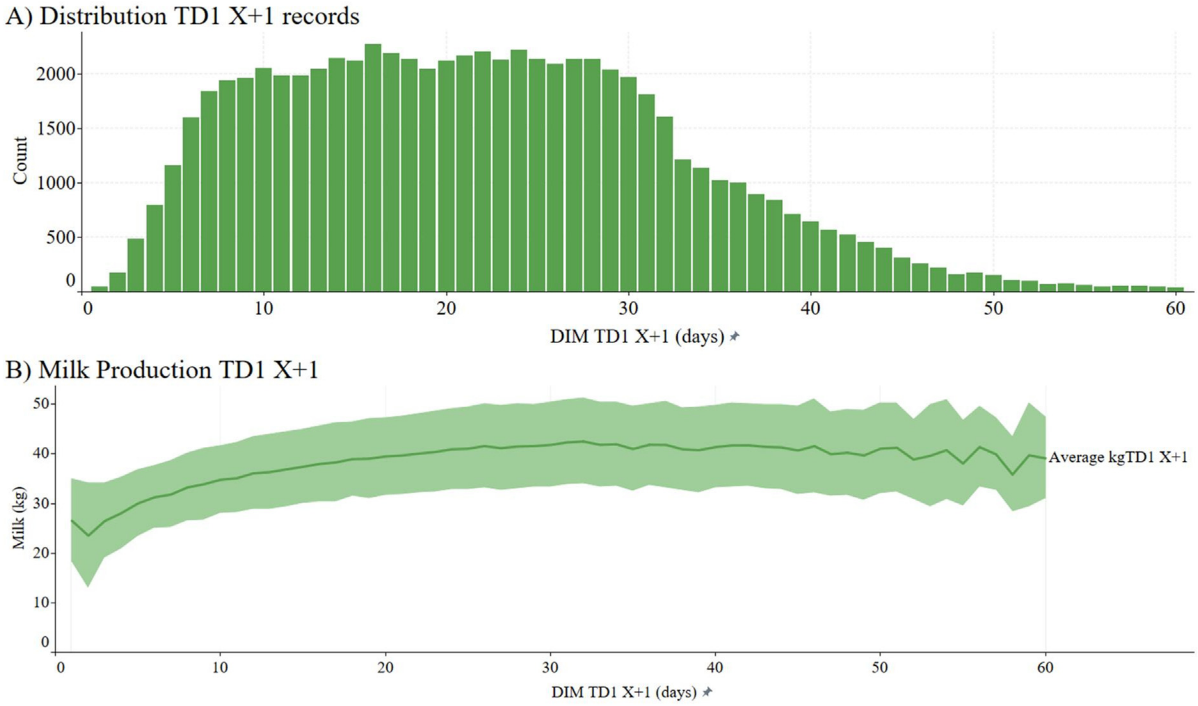

Fig. 4. Descriptive plots of TD1$_{X+1}$, in panel A the distribution of the records in function of DIM, where we can see a clear plateau in the number of cow records between 7 and 30 DIM. In panel B, an overview of the evolution of milk production during the first test day is plotted. This average reaches a plateau around days 30 DIM. Abbreviations: DIM = Days in milk, TD1$_{X+1}$ = first test day in lactation X + 1.

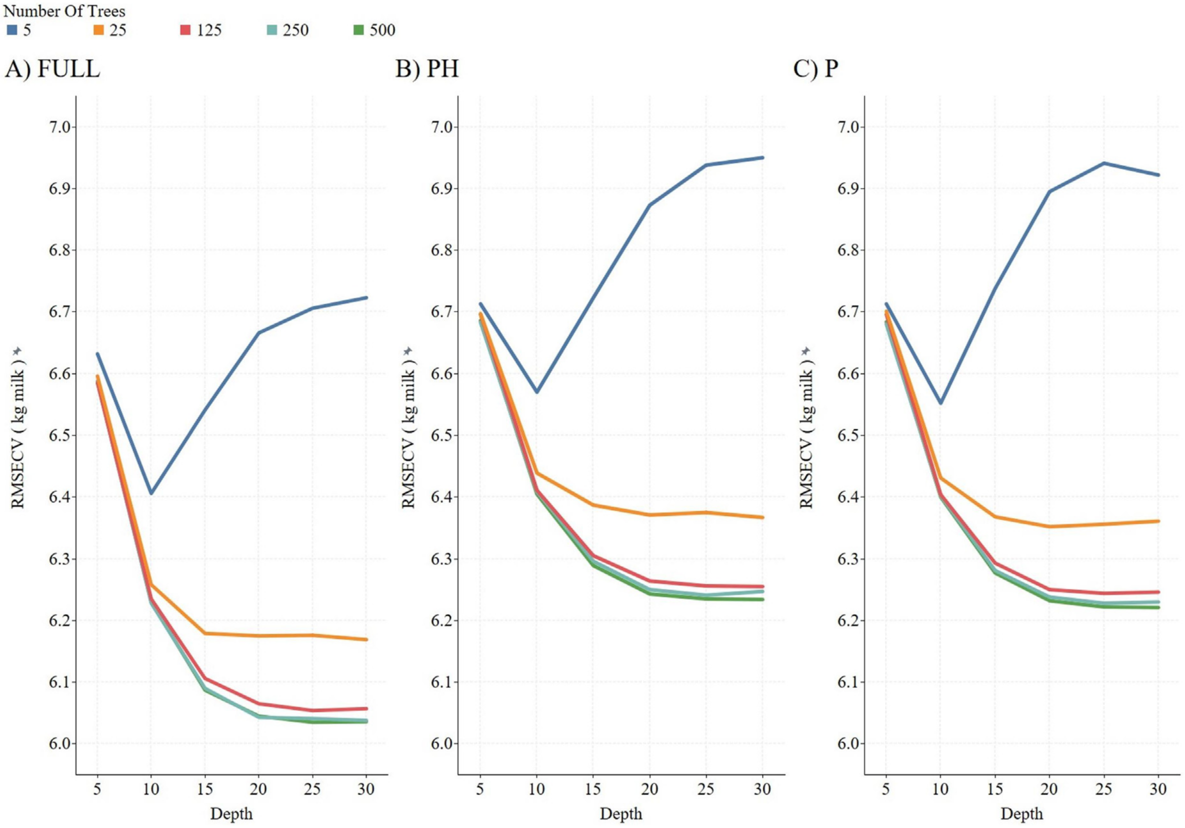

Fig. 5. The RMSECV plotted for each of the hyperparameter set for each of the three feature sets extracted from dairy cows. On the x-axis, the max depth of the trees trained in the random forest is shown. Each line representing a number of trees of those random forest models. Abbreviations: RMSECV = RMSE in cross-validation, PH = production and herd, P = production.

Table 2
Model performance indicators of the nextMILK models and benchmark models in dairy cows.

Model	RMSE	MAE	R^2	MAPE	Significant difference[1]
nextMILK$_{FULL}$	6.08	4.56	0.52	0.1327	
nextMILK$_{PH}$	6.24	4.68	0.49	0.1369	
nextMILK$_{P}$	6.18	4.58	0.51	0.1339	
Benchmark I – HERD-HERD – 75	7.48	5.64	0.27	0.1608	#, †, ¢
Benchmark I – HERD-HERD – 305	7.89	6.59	0.09	0.1807	#, †, ¢
Benchmark II – ANIMAL-HERD – 75	7.37	5.50	0.30	0.1585	#, †, ¢
Benchmark II – ANIMAL-HERD – 305	7.40	5.59	0.29	0.1608	#, †, ¢

Abbreviations: MAE = mean absolute error, MAPE = mean absolute percentage error.

[1] #, †, ¢ indicate that the squared residuals of the models were significantly higher ($P < 0.05$) when compared with respectively nextMILKFULL, nextMILKPH and nextMILKP.

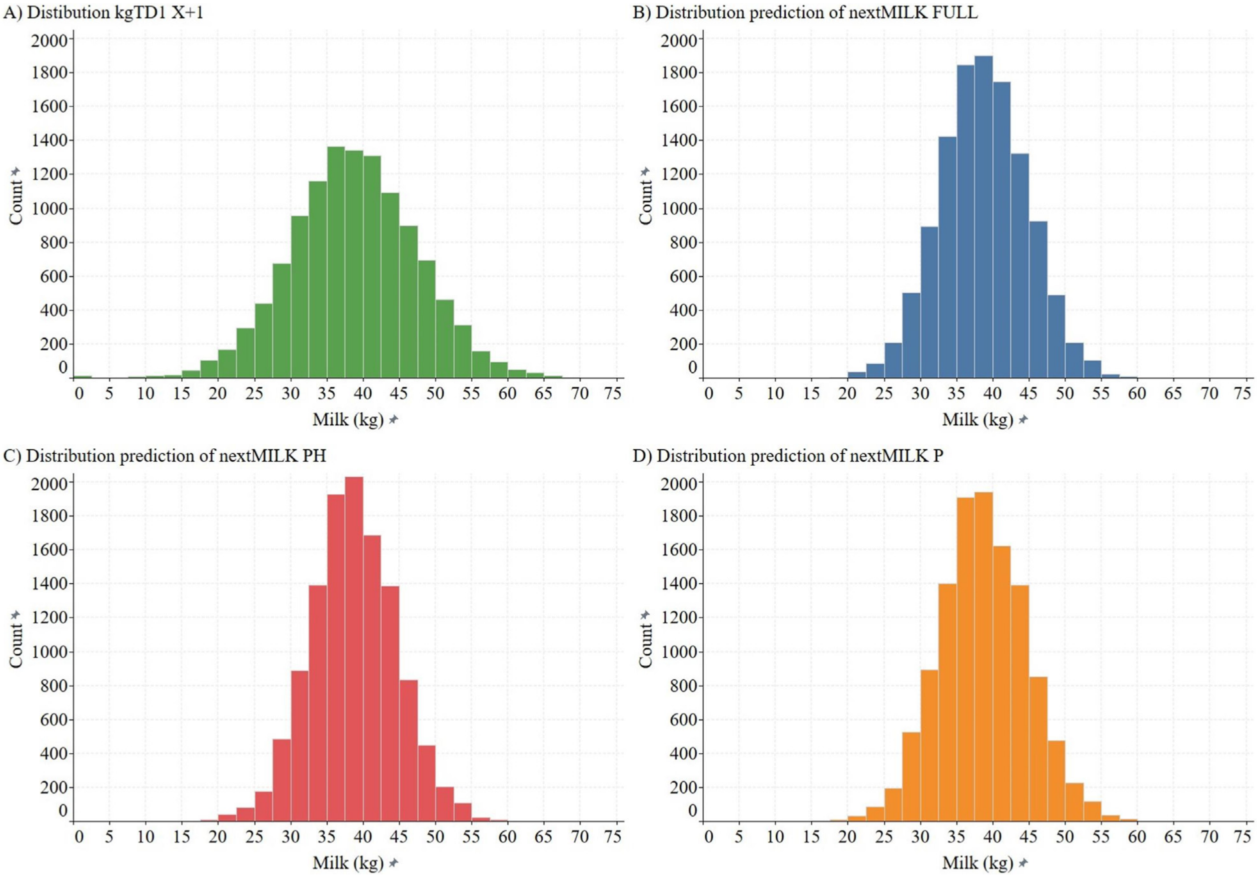

Fig. 6. Distribution of the dependent variable kg produced by the dairy cow for the complete test set in panel A, in the three other panels, the predictions of all nextMILK models are plotted for the test set. Abbreviations: TD1$_{X+1}$ = first test day in lactation X + 1, PH = production and herd, P = production.

the models are based on biological process such as the steep incline of the lactation curve at the start of lactation.

In 2012, Cole et al. described a theoretical minimum RMSE of 6 kg for predicting daily milk productions in cows, which is due to the variability resulting from changes in environment and health. The RMSE of the nextMILK models indicates that the nextMILK models are performing in the same order of magnitude as this described theoretical minimum. Furthermore, the comparison of residuals and MPI between the benchmark models and the nextMILK puts these MPIs in perspective of more simple approaches. In all cases, the nextMILK models showed significantly lower residuals and better MPI.

A possible explanation of the small difference in MPI over all the nextMILK models could be found in the feature importance vectors. The low importance of the reproductive and herd features indicates their limited contribution to the prediction, hence the absence of substantial differences between models. In terms, the comparison of these models with even smaller models should be made to evaluate from which point a real difference can be observed.

The absence of a high-quality disease registration dataset was one of the biggest limitations of the present study. It would have allowed us to select unperturbed lactations in a more objective way. Nevertheless, the applied data selection steps are set up to exclude abnormal lactations (in length, number of records) and possibly perturbed lactations. As a filtering step was applied to obtain high-quality data, it should be further investigated how this affects the model performance for an extensive dataset with qualitative disease registration.

The creation of three distinct feature sets was motivated by the variation in quality and ease of collection of the three types of features. The individual and herd-level production features are

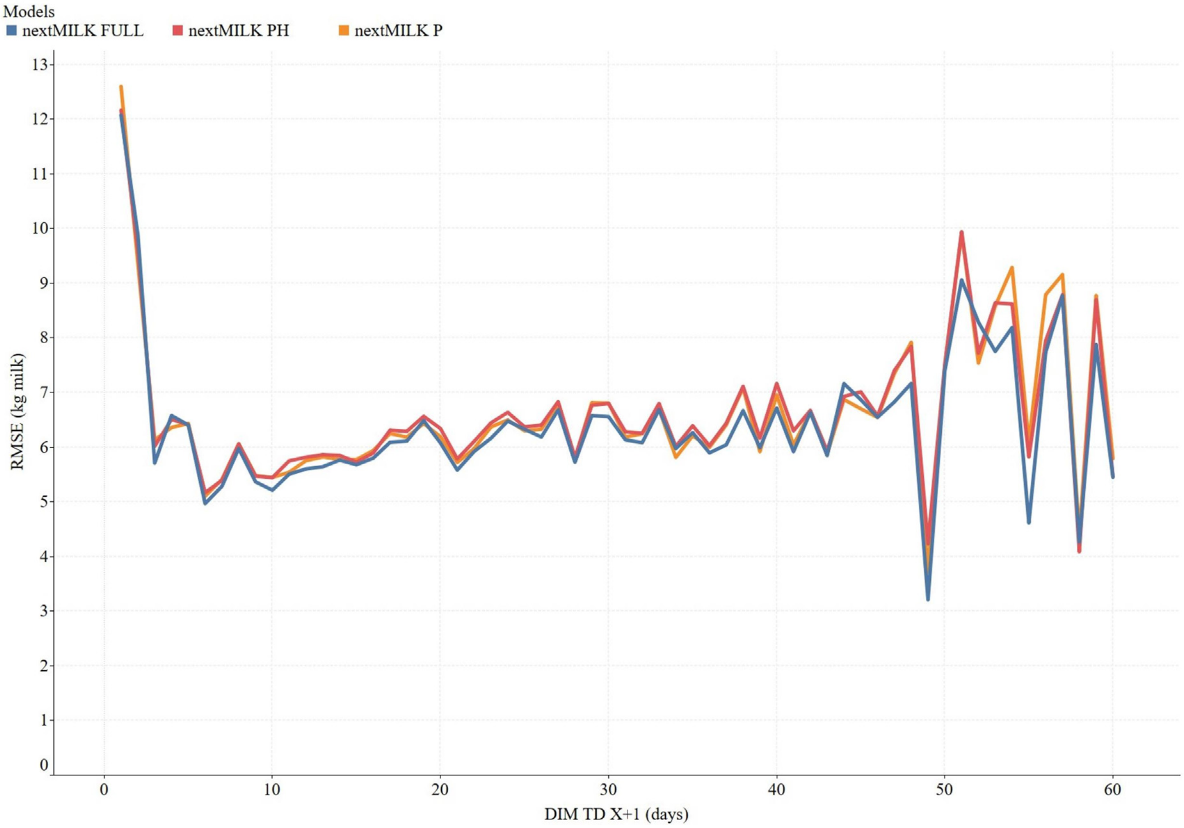

Fig. 7. In this figure, the RMSE is plotted for the test set in function of the DIM of TD1$_{X+1}$ for the nextMILK models for dairy cows. Of note: the lower number of TD1$_{X+1}$ records within the range 0–5 DIM and 40–60 DIM is displayed in Fig. 4A. Seems to be linked with a high variation of the RMSE in the same region of DIM. Abbreviations: DIM = Days in milk, TD1$_{X+1}$ = first test day in lactation X + 1, PH = production and herd, P = production.

Table 3
Top 10 feature importance extracted from the nextMILK models for dairy cows.[1]

	Importance		
Feature	FULL	PH	P
DIM TD1$_{X+1}$	0.185	0.187	0.193
M305$_X$	0.129	0.117	0.128
kgTD5$_X$	0.060	0.057	0.065
kgTD4$_X$	0.046	0.060	0.065
Milk Maximum$_X$	0.030	0.038	0.040
kgTD6$_X$	0.030	0.038	0.039
Days Dry$_X$	0.027		
Season$_{X+1}$	0.026	0.033	0.037
Δ Herd Average Days Dry$_X$	0.025		
kgTD3$_X$	0.023	0.028	0.031

Abbreviations: DIM = days in milk, X = lactation X, X+ 1 = lactation X + 1, TD1 = first test day, TD3 = third test day, TD4 = fourth test day, TD5 = fifth test day, TD6 = sixth test day, M305 = cumulative milk yield after 305 days, PH = production and herd, P = production

[1] The table with the 40 features used in the models can be found in Supplementary Table S1.

routinely collected with a high-quality standard, whereas the reproduction-derived features are considered poor in quality and challenging to collect on farms, because they often require manual inputs from the farmer.

To our current knowledge, only three studies with a comparable research question have been published. In his initial publication, Nordlund (2006) described the development of the Transition Cow Index (TCI), a tool to evaluate the success of the transition of individual animals. The TCI uses a mixed model composed of 14 parameters such as DIM on the first test day, previous 305d milk yield, lactation length of the prior lactation, SCC log score on the last test day of the prior lactation, days dry and the milking frequency of the current lactation to predict the first test day production and the 305-day milk production of the next lactation. The publication provides an extensive validation and performance of the TCI, while remaining elusive on the MPI of the model providing the TCI. This makes it impossible to compare the nextMILK models' performance with Nordlund's model.

The study of Dallago et al. (2019) focuses on predicting the production on the first test day of first lactation animals only. The authors explored three different modelling techniques: multivariate linear regression, random forest and an artificial neural network. The RMSE reported in that study ranges from 5.02 to 5.10 kg of milk, whereas their MAE ranges from 3.9 to 4.0 kg and the R^2 from 0.30 to 0.32 across the three modelling techniques. In this study, the author states that the artificial neural network model performed consistently better than the other ones. Using the MPI reported by the authors, compared to the ones of nextMILK, it seems the RMSE, the MAE and R^2 are respectively lower, lower and higher in their study. We believe that the usage of features collected on the 1st test day of the 1st lactation as inputs for the models such as %fat; % protein, SCC could be the reason for the better performances of the models presented by Dallago et al. (2019) compared to the nextMILK models and even to the theoretical minimum RMSE described by Cole et al. (2012).

In the present study, we chose to not include information of the TD we predict in the feature set, to (1) keep our predictors independent of the outcome variable, and (2) because we aim at predicting the production when no health problems influence the production. If for example composition of the predicted TD would be included, this would affect these estimations.

The SLMYP model presented by Liseune et al. (2021) predicts the 305 productions of the next lactation using daily milk meter data. Even though the forecasting horizon of this model and nextMILK models differ, the author provides MPI in function of different forecasting horizons allowing their comparison. They reported an MAE for the 0–60 DIM time window of 5.8 kg, which is higher than the MAE calculated for the nextMILK models (4.56–4.68 kg). Moreover, the nextMILK model is designed to predict the daily milk production on the first test day, between 0 and 60 days after calving, whereas the model of Liseune et al. (2021) aimed at predicting the sequence of individual daily productions for the entire 305 days, including the transition from an increasing milk production in the first part of the lactation to a decreasing milk production after peak lactation.

The nextMILK models could be used on farms as a data-driven monitoring tool providing information in the short, medium and long term. In the short term, the nextMILK models could be utilised at animal level, providing ad hoc decision support for the farmer around the transition period. In the medium term, the nextMILK models could be applied to identify the generally expected production at the start of a lactation within the farm and assess any differences over different groups (e.g. age, pens) or time. Additionally, aggregating the transition failures over time at herd level could also provide a tangible tool for the farmer with which they could assess their transition management. In the long term, the nextMILK models could be used as part of breeding programmes, to evaluate consistency over lactations or the general tendency to transition success. Although all these potential applications seem promising, intensive validation is needed to investigate the extent to which the nextMILK models could fulfil these expected goals. In particular, it is challenging to estimate the effect of animals culled before the first TD due to transition-associated-diseases on the monitoring capabilities of the nextMILK models. Furthermore, the performance increase between the nextMILK model and the benchmark may be significant, but the biological relevance of this reduction in RMSE should be investigated when validating these models.

The usage of TDR in this research on the one hand provides a wide application basis by allowing these models to be run on all farms that participate in milk recording programmes. Additionally, these programmes are a familiar resource for the farmer where a benchmark of the herd and individual performance is regularly provided, even though the exact details on the benchmark calculation in these programmes remain largely unknown and dependent on the milk recording companies. On the other hand, TDRs are intrinsically limited in views of the possible implementation mentioned hereabove. These limitations are due to the interval with which TDRs are recorded, which could cause the TDR to be recorded too late. Nevertheless, if during validation the power of predicting transition failure is proven, an altered way of recording production in the early stages of lactation could be envisaged.

In their current state, the nextMILK models do not utilise all the data traditionally being collected on a TDR such as milk fat and milk protein content, as not for all TDRs, this information was available. Not using the milk constituents allows to, in the future, use a similar modelling approach with data collected automatically by on-farm milk meters. Still, for the development of future models and when the additional information such as fat%, protein% or SCC is available and reliable, we consider it interesting to consider them as new features for the nextMILK models.

Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.animal.2022.100658.

Ethics approval

Not applicable.

Data and model availability statement

None of the data were deposited in an official repository but are available upon request.

Author ORCIDs

Matthieu Salamone: https://orcid.org/0000-0002-1505-7718
Ines Adriaens: https://orcid.org/0000-0001-9768-2308
Andy Vervaet: https://orcid.org/0000-0001-9789-487X
Geert Opsomer: https://orcid.org/0000-0002-6131-1000
Hadi Atashi: https://orcid.org/0000-0002-6853-6608
Veerle Fievez: https://orcid.org/0000-0001-5042-6200
Ben Aernouts: https://orcid.org/0000-0001-6266-3019
Miel Hostens: https://orcid.org/0000-0001-5376-976X

Author contributions

Matthieu Salamone: Conceptualisation, Methodology, Software, Formal analysis, Visualisation, Writing - Original Draft
Ines Adriaens: Methodology, Writing - Original Draft, Supervision, Writing - Review and Editing
Andy Vervaet: Writing - Review and Editing
Geert Opsomer: Writing - Review and Editing
Hadi Atashi: Methodology, Review and Editing
Veerle Fievez: Conceptualisation, Funding acquisition, Review and Editing
Ben Aernouts: Conceptualisation, Supervision, Resources
Miel Hostens: Conceptualisation, Supervision, Data Curation, Resources

Declaration of interest

The authors have not stated any conflicts of interest.

Acknowledgements

The authors would like to thank Bonny van Ranst and Kristien Neyens from MmmooOgle to allow us access to the dataset.

Financial support statement

This study is funded by VLAIO (Vlaamse Agentschap Innoveren & Ondernemen, Belgium) as part of the LA-trajectory ‘Veerkracht’, project No.170830.

References

Adriaens, I., Huybrechts, T., Aernouts, B., Geerinckx, K., Piepers, S., De Ketelaere, B., Saeys, W., 2018. Method for short-term prediction of milk yield at the quarter level to improve udder health monitoring. Journal of Dairy Science 101, 10327–10336. https://doi.org/10.3168/JDS.2018-14696.

Adriaens, I., Van Den Brulle, I., Geerinckx, K., D’Anvers, L., De Vliegher, S., Aernouts, B., 2021. Milk losses linked to mastitis treatments at dairy farms with automatic milking systems. Preventive Veterinary Medicine 194, 105420. https://doi.org/10.1016/J.PREVETMED.2021.105420.

Ben Abdelkrim, A., Tribout, T., Martin, O., Boichard, D., Ducrocq, V., Friggens, N.C., 2021. Exploring simultaneous perturbation profiles in milk yield and body weight reveals a diversity of animal responses and new opportunities to identify resilience proxies. Journal of Dairy Science 104, 459–470. https://doi.org/10.3168/jds.2020-18537.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29, 1165–1188.

Borchers, M.R., Chang, Y.M., Proudfoot, K.L., Wadsworth, B.A., Stone, A.E., Bewley, J. M., 2017. Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. Journal of Dairy Science 100, 5664–5674. https://doi.org/10.3168/jds.2016-11526.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cole, J.B., Ehrlich, J.L., Null, D.J., 2012. Short communication: Projecting milk yield using best prediction and the MilkBot lactation model. Journal of Dairy Science 95, 4041–4044. https://doi.org/10.3168/jds.2011-4905.

Dallago, G.M., Figueiredo, D.M. de, Andrade, P.C. de R., Santos, R.A. dos, Lacroix, R., Santschi, D.E., Lefebvre, D.M., 2019. Predicting first test day milk yield of dairy heifers. Computers and Electronics in Agriculture 166, 105032. https://doi.org/10.1016/j.compag.2019.105032.

De Koster, J., Salavati, M., Grelet, C., Crowe, M.A., Matthews, E., O'Flaherty, R., Opsomer, G., Foldager, L., Hostens, M., 2019a. Prediction of metabolic clusters in early-lactation dairy cows using models based on milk biomarkers. Journal of Dairy Science 102, 2631–2644. https://doi.org/10.3168/jds.2018-15533.

Dematawewa, C.M.B., Pearson, R.E., VanRaden, P.M., 2007. Modeling extended lactations of holsteins. Journal of Dairy Science 90, 3924–3936. https://doi.org/10.3168/jds.2006-790.

Ehrlich, J.L., 2011. Quantifying shape of lactation curves, and benchmark curves for common dairy breeds and parities. The Bovine Practitioner 45, 88–96.

Grelet, C., Vanlierde, A., Hostens, M., Foldager, L., Salavati, M., Ingvartsen, K.L., Crowe, M., Sorensen, M.T., Froidmont, E., Ferris, C.P., Marchitelli, C., Becker, F., Larsen, T., Carter, F., Dehareng, F., Dehareng, F., 2019. Potential of milk mid-IR spectra to predict metabolic status of cows through blood components and an innovative clustering approach. Animal 13, 649–658. https://doi.org/10.1017/S1751731118001751.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, USA.

Hermans, K., Van Ranst, B., Opsomers, G., Hostens, M., 2018. Promises and Challenges of Big Data Associated With Automated Dairy Cow Welfare Assessment. In: Butterworth, A. (Ed.), Animal Welfare in a Changing World. CABI Publishing, Wallingford, UK, pp. 199–207.

Hostens, M., Ehrlich, J., Van Ranst, B., Opsomer, G., 2012. On-farm evaluation of the effect of metabolic diseases on the shape of the lactation curve in dairy cows through the MilkBot lactation model. Journal of Dairy Science 95, 2988–3007. https://doi.org/10.3168/JDS.2011-4791.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning – with Applications in R. Springer, New York, NY, USA. https://doi.org/10.1007/978-1-4614-7138-7.

LeBlanc, S.J., 2010. Monitoring metabolic health of dairy cattle in the transition period. Journal of Reproduction and Development 56, S29–S35. https://doi.org/10.1262/jrd.1056S29.

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. Computers and Electronics in Agriculture 180, 105904. https://doi.org/10.1016/j.compag.2020.105904.

Lopreiato, V., Mezzetti, M., Cattaneo, L., Ferronato, G., Minuti, A., Trevisi, E., 2020. Role of nutraceuticals during the transition period of dairy cows: a review. Journal of Animal Science and Biotechnology 11, 96. https://doi.org/10.1186/s40104-020-00501-x.

Macciotta, N.P.P., Dimauro, C., Rassu, S.P.G., Steri, R., Pulina, G., 2011. The mathematical description of lactation curves in dairy cattle. Italian Journal of Animal Science 10, e51. https://doi.org/10.4081/ijas.2011.e51.

Mayeres, P., Stoll, J., Bormann, J., Reents, R., Gengler, N., 2004. Prediction of Daily Milk, Fat, and Protein Production by a Random Regression Test-Day Model. Journal of Dairy Science 87, 1925–1933. https://doi.org/10.3168/JDS.S0022-0302(04)73351-2.

Nordlund, K., 2006. Transition Cow Index. In: Proceedings of the 39th Annual Conference of the American Association Bovine Practitioners, 20–24 September 2006, St. Paul, MN, USA, pp. 139–143.

Parker Gaddis, K.L., Cole, J.B., Clay, J.S., Maltecca, C., 2016. Benchmarking dairy herd health status using routinely recorded herd summary data. Journal of Dairy Science 99, 1298–1314. https://doi.org/10.3168/jds.2015-9840.

Pascottini, O.B., Leroy, J.L.M.R., Opsomer, G., 2020. Metabolic stress in the transition period of dairy cows: Focusing on the prepartum period. Animals 10, 1419. https://doi.org/10.3390/ani10081419.

Poppe, M., Veerkamp, R.F., van Pelt, M.L., Mulder, H.A., 2020. Exploration of variance, autocorrelation, and skewness of deviations from lactation curves as resilience indicators for breeding. Journal of Dairy Science 103, 1667–1684. https://doi.org/10.3168/jds.2019-17290.

Probo, M., Pascottini, O.B., LeBlanc, S., Opsomer, G., Hostens, M., 2018. Association between metabolic diseases and the culling risk of high-yielding dairy cows in a transition management facility using survival and decision tree analysis. Journal of Dairy Science 101, 9419–9429. https://doi.org/10.3168/JDS.2018-14422.

Saun, V., Robert, J., 2006. Metabolic profiles for evaluation of the transition period. In: Proceedings of the 39th Annual Conference of the American Association Bovine Practitioners, 20–24 September 2006 St. Paul, MN, USA, pp. 130–138.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., Weigel, K., 2014. Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. Journal of Dairy Science 97, 731–742. https://doi.org/10.3168/jds.2013-6693.

van der Heide, E.M.M., Veerkamp, R.F., van Pelt, M.L., Kamphuis, C., Athanasiadis, I., Ducro, B.J., 2019. Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. Journal of Dairy Science 102, 9409–9421. https://doi.org/10.3168/jds.2019-16295.

Walsh, R.B., Walton, J.S., Kelton, D.F., LeBlanc, S.J., Leslie, K.E., Duffield, T.F., 2007. The effect of subclinical ketosis in early lactation on reproductive performance of postpartum dairy cows. Journal of Dairy Science 90, 2788–2796. https://doi.org/10.3168/jds.2006-560.