Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health?

S. Franceschini,[1]* C. Grelet,[2] J. Leblois,[3] N. Gengler,[1] GplusE consortium,† and H. Soyeurt[1]
[1]University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium
[2]Walloon Agricultural Research Center (CRA-W), 5030 Gembloux, Belgium
[3]Walloon Breeders Association Group (Elevéo by Awé groupe), 5590 Ciney, Belgium

ABSTRACT

Among the dairy sector's current concerns, the assessment of global animal health status is a complex challenge. Its multidimensionality means that global monitoring tools are rarely considered. Instead, specific disease detection is often studied separately and, due to financial and ethical issues, uses small-scale data sets focusing on few biomarkers. Several studies have already been conducted using milk Fourier transform mid-infrared (FT-MIR) spectroscopy to detect mastitis and lameness or to quantify health-related biomarkers in milk or blood. Those studies are relevant but they focus mainly on one biomarker or disease. To solve this issue and the small-scale data set, in this study, we proposed a holistic approach using big data obtained from milk recording, including milk yield, somatic cell count, and 27 FT-MIR–based predictors related to milk composition and animal health status. Using 740,454 records collected from 114,536 first-parity Holstein cows in southern Belgium, we performed repeated unsupervised learning algorithms based on Ward's agglomerative hierarchical clustering method to find potential interesting patterns. A divide-and-conquer approach was used to overcome the limitation of computational resources in clustering a relatively large data set. Five groups of records were identified. Differences observed in the fourth group suggested a relationship to metabolic disorders. The fifth group seemed to be related to mastitis. In a second step, we performed a partial least squares discriminant analysis (PLS-DA) to predict the probability of belonging to those specific groups for the entire data set. The obtained global accuracy was 0.77 and the balanced accuracy (i.e., the mean between sensitivity and specificity) of discriminating the fourth and fifth groups was 0.88 and 0.96, respectively. Then, a validation of the interpretation of those groups was performed using 204 milk and blood reference records. The predicted probability associated with the metabolic disorders issue had significant correlations of 0.54 with blood β-hydroxybutyrate, 0.44 with blood nonesterified fatty acids, −0.32 with blood glucose, −0.23 with milk glucose-6-phosphate, and 0.38 with milk isocitrate. In contrast, the predicted probability of belonging to the mastitis group had correlations of 0.69 with milk lactate dehydrogenase, 0.46 with milk N-acetyl-β-D-glucosaminidase, −0.18 with milk free glucose, and 0.16 with milk glucose-6-phosphate. Consequently, these results suggest that the obtained quantitative traits indirectly reflect some of the main health disorders in dairy farming and could be used to monitor dairy cows on a large scale. By using unsupervised learning on large-scale milk recording data and then validating the pattern using reference laboratory measures, we propose a new approach to quickly assess dairy cow health status.

Key words: big data, animal health, unsupervised learning, milk, mid-infrared

INTRODUCTION

Monitoring dairy cow health is complex because it involves many aspects such as disease, energy balance, and heat stress. However, dairy cow health, being a part of animal welfare, is important for dairy farming sustainability. Stakeholders (producers, policymakers, and consumers) need warranties about the ethics of the

product. In particular, the welfare status of cows is of increasing concern (Fraser, 1995; Blokhuis et al., 2003).

Currently, cow health monitoring tools mostly rely on milk and blood analyses, which provide health-related biomarkers. However, defining health indicators is a lengthy and costly process because it requires many potential biomarkers to be scanned among an animal population that has health issues (Foldager et al., 2020; Krogh et al., 2020). Depending on the issue, the process can raise ethical questions when the disorder is induced in animals or the measurement is invasive. In other cases, detecting animals with health issues is complicated because the farms participating in the studies are already aware of the problem, making the sample unrepresentative and biased toward healthy animals.

Animal scientists have developed strategies to predict health indicators using milk Fourier transform mid-infrared (**FT-MIR**) spectrometry to avoid ethical issues and the cost of routine measurements of biomarkers. This spectroscopic technology has many advantages as it is noninvasive and reflects part of the animal's metabolism by assessing the global milk composition. Moreover, milk FT-MIR spectra, along with other dairy production phenotypes, are collected by DHI organizations on average once a month in participating herds. From the FT-MIR spectra, statistical models can predict biomarkers of interest such as acetone, BHB, citrate, lactoferrin, and SCC to assess and, therefore, monitor dairy cow health (Soyeurt et al., 2012; Luke et al., 2019). This technique enables routine individual assessment of cow health status on farms after applying appropriate prediction equations related to the biomarkers to the recorded spectral data.

A standard procedure to obtain prediction models consists of collecting blood and milk samples from as many healthy and sick animals as possible, analyzing those milk samples using FT-MIR spectrometry to record the spectral data, and measuring the contents of the studied biomarkers in milk and blood using certified chemical methodologies. Afterward, supervised machine learning methods such as partial least squares (**PLS**) regression can be used to build the model, predicting the desired biomarker (Soyeurt et al., 2020). This process has been largely used in the last decades to increase the number of phenotypes predicted by FT-MIR and not only to predict health issues (De Marchi et al., 2014; Bastin et al., 2016; Gengler et al., 2016). Table 1 presents the characteristics of some models published in the literature predicting different health biomarkers naturally present in milk and blood. A large range of cross-validation accuracy can be observed: R^2 varied between 0.21 and 0.88 (Table 1). More information on key factors affecting the quality of predictions is available in Grelet et al. (2021). Model performance

and validation accuracy can sometimes be poorer for several reasons. First, for practical, ethical, or financial issues, the calibration data set may not be representative enough, with too few animals or herds. However, to maximize the prediction quality, the modeling procedure requires sampling among the whole animal population to maximize the variability of the calibration data used to develop the equation (Grelet et al., 2021). Thus, making a prediction outside of the calibration variability is detrimental to the prediction quality due to inner extrapolation. In addition, maximizing the calibration set for health indicators requires both healthy and sick animals to be included. Sick animals are relatively scarce, which explains why obtaining high BHB levels is more difficult, as mentioned by Grelet et al. (2016), Fleming et al. (2017), and Bonfatti et al. (2019). This unbalanced situation between sick and healthy animals can affect the final prediction accuracy of the developed models. Finally, most of the metabolites of interest have a very low concentration in milk or are even predicted in another matrix (e.g., blood). These indirect predictions rely more on global milk changes than on measuring a specific molecule. Consequently, the disorder cannot be simulated artificially, such as spiking milk with molecules of interest, because it would interfere with the natural relationships existing between the milk components. All these aspects explain the difficulty in developing prediction models related to health indicators. To solve these issues, it would be interesting to combine different health biomarker predictions, different models derived from different populations, and usual production traits and indicators to provide insights into cow health status (e.g., milk production or BW).

In their review of infrared spectrometry as a high-throughput phenotyping technology, Bresolin and Dórea (2020) suggest using larger data sets and modern data-mining approaches to improve the models. In the current study, the innovative aspect consists of realizing an unsupervised learning approach that profits from the large-scale data set collected from DHI and combining several FT-MIR–based predictions and usual dairy traits to discriminate groups of cows that have a different data pattern. This approach allows the combination of several indicators of health issues (i.e., multidimensional aspect). Furthermore, the exploration of a large DHI database ensures the consideration of high variability and warrants the representativeness of the sample set. Therefore, the context of this study can be considered as a "big data" case; the 5 Vs representing aspects of big data are found: volume, variety, veracity, velocity, and value (Gandomi and Haider, 2020).

Moreover, unsupervised methods have the advantage of not introducing any bias about the definition of

Table 1. Published prediction models for health biomarkers

Trait	Unit	No. of records	No. of cows	No. of herds	R^2cv[1]	Reference
Energy balance	MJ	1,535	378	1	0.67–0.78	McParland and Berry, 2016
	MJ/d	11,525–30,429	1,101	1	0.42–0.77	Smith et al., 2019
Energy intake	MJ/d	11,940–30,565	1,101	1	0.39–0.75	Smith et al., 2019
	MJ	1,535	378	1	0.64–0.88	McParland and Berry, 2016
Blood BHB	mmol/L	1,914	826	55	0.21–0.43	Belay et al., 2017
	mmol/L	3,629	1,013	3	0.27–0.46	Pralle et al., 2018
	mmol/L	380	241	—	0.70	Grelet et al., 2019
	mmol/L	1,910	542	2	0.50–0.58	Bonfatti et al., 2019
Blood glucose	mmol/L	380	241	—	0.44	Grelet et al., 2019
Blood IGF-1	mg/L	387	241	—	0.61	Grelet et al., 2019
Blood NEFA[2]	uEq/L	380	241	—	0.39	Grelet et al., 2019
Milk lactoferrin	mg/L	2,499	—	>9	0.71	Soyeurt et al., 2012
	mg/L	5,541	—	>9	0.60	Soyeurt et al., 2020

[1]R^2cv = accuracy of cross-validation.

[2]Nonesterified fatty acids.

health. Among those methods, clustering, and particularly hierarchical clustering, is used as it allows patterns to be observed at any scale of the database and also cuts the data into groups. Previous studies have considered multiple reference health measures together and even clustering, but they were conducted using small-scale reference data sets (LeBlanc, 2010; Grelet et al., 2019; Atashi et al., 2020) and not from a routine DHI database. Therefore, the current approach presents the distinction of having no reference analysis of health biomarkers to construct the discrimination between healthy and sick animals. However, this discrimination will be validated using real reference measurements to attest to the relevancy of the discovered groups.

MATERIALS AND METHODS

All data treatments and analyses were performed using R software (version 4.0.2; https://www.r-project .org/). All data were collected as per standard routine by the Walloon Breeders Association, which is accredited to perform such collection.

Calibration Data

The data set to be used with unsupervised learning methods, hereafter called the calibration data set, contained 740,454 DHI records collected by Elevéo (Awé groupe, Ciney, Belgium) in 1,077 herds located in southern Belgium between January 2012 and March 2020 from 114,536 primiparous Holstein cows with DIM ranging between 5 and 365 d. Milk analysis was performed by the milk laboratory Comité du Lait (Battice, Belgium) on Foss Milkoscan FT6000, FT+, and FT7 standardized spectrometers. Among the large collection of possible phenotypes for this study, we kept only milk yield, milk FT-MIR spectra, SCC, and contents of fat, protein, urea, and lactose predicted by the spectrometer. Furthermore, predictive equations developed from previous studies (Table 2) were applied to the recorded spectral data to predict 20 additional biomarkers related to health issues. Although initially predicted in grams per deciliter (g/dL) of milk, the contents of SFA, C18:1 *cis*-9, MUFA, and PUFA were expressed in grams per 100 grams of fat using the fat content predicted by the spectrometer, as this unit is more relevant for assessing the biological pathways behind fatty acid production. Then, DMI and fat- and protein-corrected milk (**FPCM**) were computed using the following equations (NRC, 2001; Yan et al., 2011):

$$DMI = \left(0.372FPCM + 0.0968BW^{0.75}\right)$$
$$\times \left[1 - \exp\left(-0.192WOL - 0.70464\right)\right] \text{ and}$$

$$FPCM = MY\left[0.337 + \left(0.116FAT\right)\right] + 0.06PROT,$$

where BW is the predicted BW using the equation of Tedde et al. (2021), WOL is the week of lactation, MY is milk yield, and FAT and $PROT$ are fat and protein contents expressed as percentages. The consumption index was defined as the ratio of DMI to FPCM. Finally, the calibration data set used for the clustering contained 29 traits that were directly or indirectly related to health issues (Table 3).

Clustering Method

Patterns were found in the calibration data set using hierarchical clustering. This agglomerative algorithm works step by step to combine the closest records or groups of records until the whole data set is considered. Ward's agglomerative hierarchical clustering method

Table 2. Performance of the equations used[1]

Equation[2]	Unit	N	R^2c	R^2cv	SEc	SEcv	RPDcv	Reference
BW	kg	721	0.51	0.47	51	53	1.37	Tedde et al., 2021
CH_4	g/d	1,089	0.73	0.68	53	57	1.78	Vanlierde et al., 2021
Sodium	mg/kg of milk	1,019	0.48	0.44	49.25	50.98	1.34	Christophe et al., 2021
Calcium	mg/kg of milk	1,094	0.83	0.82	51.79	53.38	2.34	Christophe et al., 2021
Phosphorus	mg/kg of milk	1,083	0.76	0.75	57.36	58.71	1.99	Christophe et al., 2021
Magnesium	mg/kg of milk	1,124	0.73	0.72	6.35	6.53	1.88	Christophe et al., 2021
Potassium	mg/kg of milk	1,090	0.57	0.55	85.79	88.14	1.48	Christophe et al., 2021
Lactoferrin	mg/L	924	0.74	0.71	125	131	1.86	Soyeurt et al., 2012
EB	%	1,010	0.48	0.43	4.84	5.04	1.33	Grelet et al., 2017
DMI2	kg/d			0.71		5.70	1.83	Grelet et al., 2021
NUE	%	1,093	0.55	0.52	3.05	3.17	1.44	Grelet et al., 2020
Citrate	mmol/L	498	0.90	0.89	0.68	0.72	3.04	Grelet et al., 2016
Acetone	mmol/L	201	0.69	0.62	0.19	0.21	1.63	Grelet et al., 2016
BHB	μmol/L	419	0.75	0.71	0.13	0.14	1.85	Grelet et al., 2016
RFI	kg/d	1,008	0.52	0.48	2.73	2.87	1.39	Grelet et al., 2017
RFI2	kg/d	1,013	0.46	0.43	2.77	2.89	1.31	Grelet et al., 2017
SFA	g/dL of milk	1,790	0.99	0.99	0.07	0.07	10.22	Grelet et al., 2014
C18:1 *cis*-9	g/dL of milk	1,781	0.95	0.95	0.06	0.06	4.35	Grelet et al., 2014
MUFA	g/dL of milk	1,793	0.97	0.97	0.06	0.06	5.83	Grelet et al., 2014
PUFA	g/dL of milk	1,799	0.78	0.77	0.02	0.02	2.10	Grelet et al., 2014

[1]R^2c = coefficient of determination of calibration; R^2cv = coefficient of determination of cross validation; SEc = standard error in calibration; SEcv = standard error in cross validation; RPDcv = ratio of performance to deviation in cross validation.

[2]EB = energy balance; DMI2 = predicted DMI; NUE = nitrogen use efficiency; RFI = residual feed intake (alternate calculation via RFI2); C18:1 *cis*-9 = oleic acid.

with Euclidean distance was chosen because this is widely used to distinguish groups in a multivariate Euclidean space (Ward, 1963). However, this type of clustering has the constraint of requiring the computing of a distance matrix between every observation or cluster at each step. So, for n records, the algorithm computes $n(n-1)/2$ distances. For 740,454 records, this represents 274×10^9 distances, which would make the computation too slow, and storing the distance matrix in memory would be impossible with standard computational infrastructure. The volume of data is a common hindrance in big data methodologies, and many approaches exist to solve this issue (Wang et al., 2016). In this study, we decided to apply a divide-and-conquer approach, which consisted of dividing the data into more manageable subsets, analyzing them separately, and recombining them afterward. The calibration data were first split randomly into 100 subsets of ~7,500 records. For each subset, the records were clustered, and the obtained dendrogram was cut to divide the subset into 500 groups. A mean sample (i.e., a centroid) was created for each group by averaging all features. Therefore, 50,000 centroids (i.e., 500 centroids from 100 subsets) were computed to summarize the information contained in the initial data set. Those centroids were then clustered using the same hierarchical clustering method. As the approach depends on the splitting of the data set, the procedure was replicated 10 times to estimate the robustness of the final clustering obtained. From this final clustering, the inertia (i.e., the ratio

between the between-class and within-class variance) of the dendrogram was estimated to define the best number of groups (i.e., selection of several groups before a large drop of inertia gain). This can be done by looking at the graph of the inertia according to the number of groups.

Cluster Interpretation

Standard statistical methods to test the significance of the effects, such as ANOVA, are useless here because of the number of records. There are too many observations, so the tests are too powerful, making every effect significant. Moreover, the main objective was to interpret the clusters to understand what they represent. Thus, the differences between clusters were represented by calculating their least squares means (**LSM**) per group, which were visualized using bar plots on a standardized scale.

Expanding the Result to the Original Data Set

For the previous step, the large data set was simplified into 50,000 centroids clustered into several groups. To expand this result to the whole data set, a cluster prediction model was needed. A PLS discriminant analysis (**PLS-DA**) model was built from the centroids using the 29 biomarkers. It is a predicting discriminant model using linear combinations of the original variables called latent variables, which are optimized to best represent

Table 3. Descriptive statistics of the mid-infrared–based predictors, milk yield, and SCC from the calibration data set (n = 740,454 records)

Abbreviation[1]	Unit	Minimum	Maximum	Mean	SD	CV
Milk	kg/d	8.10	78.00	23.72	5.91	0.25
pFAT	g/dL of milk	21.10	69.90	39.66	5.99	0.15
pPROT	g/dL of milk	20.30	51.50	33.59	3.27	0.10
Lactose	g/dL of milk	3.53	5.68	4.84	0.17	0.03
Urea	mmol/L	10.00	1,260.00	249.32	79.66	0.32
SCC	cells × 1,000/mL	10.00	3,490.00	139.60	297.49	2.13
Weight	kg	401.49	934.12	620.16	65.20	0.11
CH_4	g/d	100.00	699.99	306.31	58.38	0.19
Na	mg/kg of milk	200.02	499.98	321.90	34.52	0.11
Ca	mg/kg of milk	900.00	1,500.00	1,154.47	97.32	0.08
P	mg/kg of milk	700.62	1,299.95	997.53	83.77	0.08
Mg	mg/kg of milk	80.00	130.00	100.68	7.73	0.08
K	mg/kg of milk	1,100.03	1,700.00	1,494.43	90.87	0.06
Lactoferrin	mg/L	35.11	800.34	162.11	88.15	0.54
DMI2	kg/d	14.00	33.00	24.50	3.20	0.13
DMI	kg/d	7.65	38.51	19.85	2.67	0.13
EB	%	−19.38	10.00	0.33	3.54	10.89
NUE	%	−2.37	27.10	11.34	2.14	0.19
Citrate	mmol/L	2.84	17.00	8.64	1.36	0.16
Acetone	mmol/L	0.01	0.30	0.06	0.03	0.46
BHB	μmol/L	38.20	399.82	158.07	42.24	0.27
RFI	kg/d	−16.09	1.50	−3.46	2.18	−0.63
RFI2	kg/d	−13.44	5.68	−2.10	2.02	−0.96
FPCM	kg/d	6.50	74.25	23.52	5.53	0.23
CI	kg of DMI/kg of milk	0.36	2.24	0.88	0.18	0.21
SFA	g/100 g of fat	30.50	178.75	70.39	4.66	0.07
C18:1 *cis*-9	g/100 g of fat	7.67	47.63	20.47	3.68	0.18
MUFA	g/100 g of fat	10.39	66.65	28.17	4.21	0.15
PUFA	g/100 g of fat	1.26	8.81	3.70	0.57	0.15

[1]pFAT = fat content; pPROT = protein content; DMI2 = predicted DMI; EB = energy balance; NUE = nitrogen use efficiency; RFI = residual feed intake (alternate calculation via RFI2); FPCM = fat- and protein-corrected milk; CI = consumption index; C18:1 *cis*-9 = oleic acid.

the variability of the response and the original variables. The output of the model is the probability of belonging to a specific cluster. The hyperparameter of the model is the number of latent variables, which was fixed based on the performances obtained from a 10-fold cross-validation. As our aim was to develop a model able to predict every cluster, but more specifically the clusters associated with sick animals, we compared the global accuracy and the balanced accuracy for each cluster, which is the mean between sensitivity and specificity. The PLS-DA model is sensitive to unbalanced data as it will most likely predict the most observed cluster. In a case where the proportions of observations in each cluster vary widely, it is recommended to generate a balanced training data set. The considered up-sampling approach consists of sampling with replacement of the original data set but with the constraint to reach 10,000 records for each cluster. Moreover, we expected that a quantitative trait would better represent the continuum of the animal health status instead of a binary division (healthy vs. sick). Compared with the *k*-nearest neighbors algorithm, PLS-DA gives more quantitative information: the scores on specific latent variables and the probability of being in a cluster.

Validation

The innovative aspect of our approach was to use a large-scale database to find patterns potentially related to animal health to keep the maximum variability existing in the population. Therefore, no reference measurements for health status were used. This is why validation was performed using reference biomarkers measured in milk and blood samples. This approach aimed to validate the interest of obtained clusters and the potential of an associated quantitative trait such as a PLS-DA probability of belonging to a cluster.

The validation health data set was created from the European Union FP7 Genotype plus Environment project (http://www.gpluse.eu). The data set contained 380 records collected between 2013 and 2018 for primiparous and multiparous Holsteins with DIM between 10 and 50. As the clusters are defined from primiparous Holsteins, a first data set containing only primiparous cows was created but only 50 records were usable as SCC measures were missing. Therefore, a second data set containing 204 records from primiparous and multiparous cows was developed to increase the number of observations available. Several biomarkers

Table 4. Descriptive statistics for the reference measurements of health-related traits from the validation sets

Measure[1]	Unit	Lactation 1 (n = 50)				All lactations (n = 204)			
		Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
Blood									
Cholesterol	g/L	1.64	5.57	3.22	0.89	1.64	7.33	3.70	1.12
Fructosamine	μmol/L	192.83	296.07	263.71	19.81	192.83	298.77	259.62	18.35
BHB	mmol/L	0.24	2.14	0.57	0.40	0.21	4.29	0.63	0.48
LogBHB	log(mmol/L)	−0.62	0.33	−0.31	0.22	−0.68	0.63	−0.27	0.22
NEFA	μEq/L	95.77	1,884.36	518.48	422.37	26.07	2,757.01	606.17	424.85
P4	ng/mL	0.12	11.07	1.89	2.23	0.02	12.92	2.48	2.80
Glucose	mmol/L	2.79	4.70	3.89	0.39	2.02	4.7	3.61	0.41
Milk									
Glu6P	mmol/L	0.05	0.44	0.20	0.09	0.05	0.44	0.17	0.08
GluFree	mmol/L	0.08	0.41	0.23	0.09	0.03	0.58	0.23	0.10
BHB	μmol/L	20.00	241.00	61.00	44.81	20	735	65.33	60.97
Isocitrate	mmol/L	0.09	0.32	0.19	0.05	0.07	0.32	0.17	0.05
Urea	mmol/L	0.95	6.80	3.37	1.46	0.47	7.17	3.11	1.38
NAGase	units/L	0.56	6.67	2.18	1.18	0.56	10.78	1.92	1.08
LDH	units/L	0.78	8.02	3.27	1.68	0.64	39.88	3.10	3.23

[1]NEFA = nonesterified fatty acid; P4 = progesterone; Glu6P = glucose-6-phosphate; GluFree = free glucose; NAGase = *N*-acetyl-β-D-glucosaminidase; LDH = lactate dehydrogenase.

were measured in individual milk and blood samples: cholesterol, fructosamine, BHB, nonesterified fatty acid (**NEFA**), progesterone (**P4**), and glucose in blood and glucose-6-phosphate, free glucose, isocitrate, urea, *N*-acetyl-β-D-glucosaminidase (**NAGase**), and lactate dehydrogenase (**LDH**) in milk (Table 4). De Koster et al. (2019) describe the data collection and the applied protocol in more detail. All phenotypes included in the calibration data set were also collected to allow prediction of the clusters, PLS-DA probabilities, and scores on latent variables for those observations.

To validate the relationships between the probabilities or the scores on latent variables and the reference values measured in the laboratory, Pearson correlations were calculated between those quantitative and reference traits. Then, statistical *t*-tests for association between paired samples were performed to assess the significance of the estimated correlation values.

RESULTS AND DISCUSSION

Developing the Clusters

As mentioned previously, standard hierarchical clustering methodologies could not be used and a specific approach was developed. As this approach involved randomness, the process was repeated 10 times to assess its robustness. To determine the similarity, we looked at whether the groups obtained and the division order between them were equivalent. Nine dendrograms out of 10 repetitions were similar. Thus, only one representative dendrogram is shown in Figure 1. Based on inertia, the dendrogram was cut into 5 groups. The

graph is structured into 2 large groups divided into one sub-group containing cluster 1 and cluster 4 and another sub-group divided into 3 sub-groups (clusters 2, 3, and 5; Figure 1). Among these 3 sub-groups, cluster 5 was very small and well separated from the others. The distribution of the centroids among the clusters was uneven, with 15,271 centroids in cluster 1 (30.54%), 19,573 in cluster 2 (39.15%), 5,452 in cluster 3 (10.90%), 6,856 in cluster 4 (13.71%), and 2,848 in cluster 5 (5.70%). Clusters 1 and 2 covered most of the records, with 69.69% of observations.

As mentioned earlier, testing for the significance of the effect is irrelevant here. Instead, the LSM of the 29 studied traits were computed and are illustrated in Figure 2 on a standardized scale. From the visual analysis, 2 clusters differed from the others: cluster 4 and cluster 5. Clusters 1 and 4 had similar behavior, with both means of each prediction being positive or negative. However, the means of cluster 4 were larger than those of cluster 1, suggesting that clusters 1 and 4 represented the same biological phenomena but with severe cases being in cluster 4. Those groups were characterized by negative values for the predictions of protein content, BW, CH$_4$, P, DMI, energy balance (**EB**), residual feed intake, consumption index, and SFA. There were positive means for the predictions of Na, nitrogen use efficiency, citrate, acetone, BHB, and the predicted percentages of C18:1 *cis*-9, MUFA, and PUFA in milk fat. Based on those results, we hypothesized that cluster 4 referred to cows with metabolic disorders induced by negative EB, up to ketosis. Although we mainly used predictions in this study, the results are similar to those of Gross et al. (2011), who observed reduced milk production
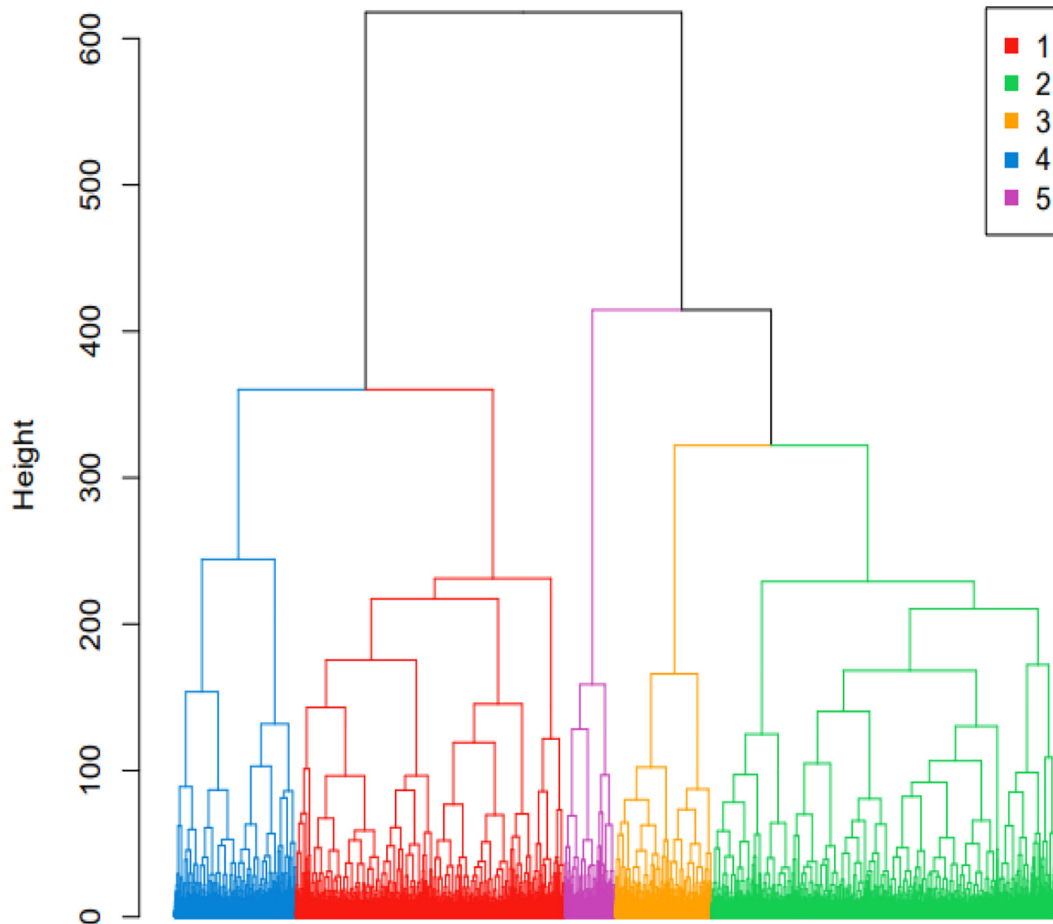
Figure 1. Dendrograms for the 50,000 centroids. Height indicates the value of the Ward's criterion for each agglomeration of centroids or groups of centroids. Agglomerations with low heights represent centroids, or groups of centroids, that are close. Clusters are identified by color.

with a lower percentage of protein, combined with a lower BW and DMI for Holstein cows in negative EB. Moreover, cluster 4 was related to lower content of SFA and higher contents of MUFA and PUFA. Vranković et al. (2017) and Fiore et al. (2020) showed that similar changes in milk fatty acid composition, with decrease of de novo fatty acids, are associated with negative EB and body fat mobilization.

Cluster 5 was characterized mostly by a large positive value for SCC. Moreover, Na and lactoferrin means were high, and that of lactose was low, suggesting that cluster 5 was related to mastitis issues (Figure 2). Although we used predictions for Na and lactose, these findings can be corroborated by previous works. Bansal et al. (2005) made the same observation for cows with subclinical mastitis using reference measurements. Fernando et al. (1985) also mentioned the use of Na as an indicator of subclinical mastitis. Lactose as an indicator of subclinical mastitis was identified by Ebrahimie et al. (2018). Cluster 3 was defined by low production

of milk associated with high estimated means for fat, protein, and minerals.

Predicting the Cluster for the Calibration Data Set

At this stage, we discovered 2 potentially interesting groups for disease detection (i.e., cluster 4 and cluster 5). However, those clusters were only defined using the centroids. Therefore, we needed to expand the results of the clustering to the whole data set. As each of the centroids was defined from 500 groups for each of the 100 subsets, this could be performed by recovering the records for each centroid. However, this method prevents the extrapolation of the results to future records. Instead, a PLS-DA model was built to allow discrimination of the found clusters from the 50,000 centroids. As the number of centroids by cluster was greatly unbalanced, which risks impairing the predictions, we used an up-sampling approach based on bootstrapping, wherein we sampled with higher probabilities for the
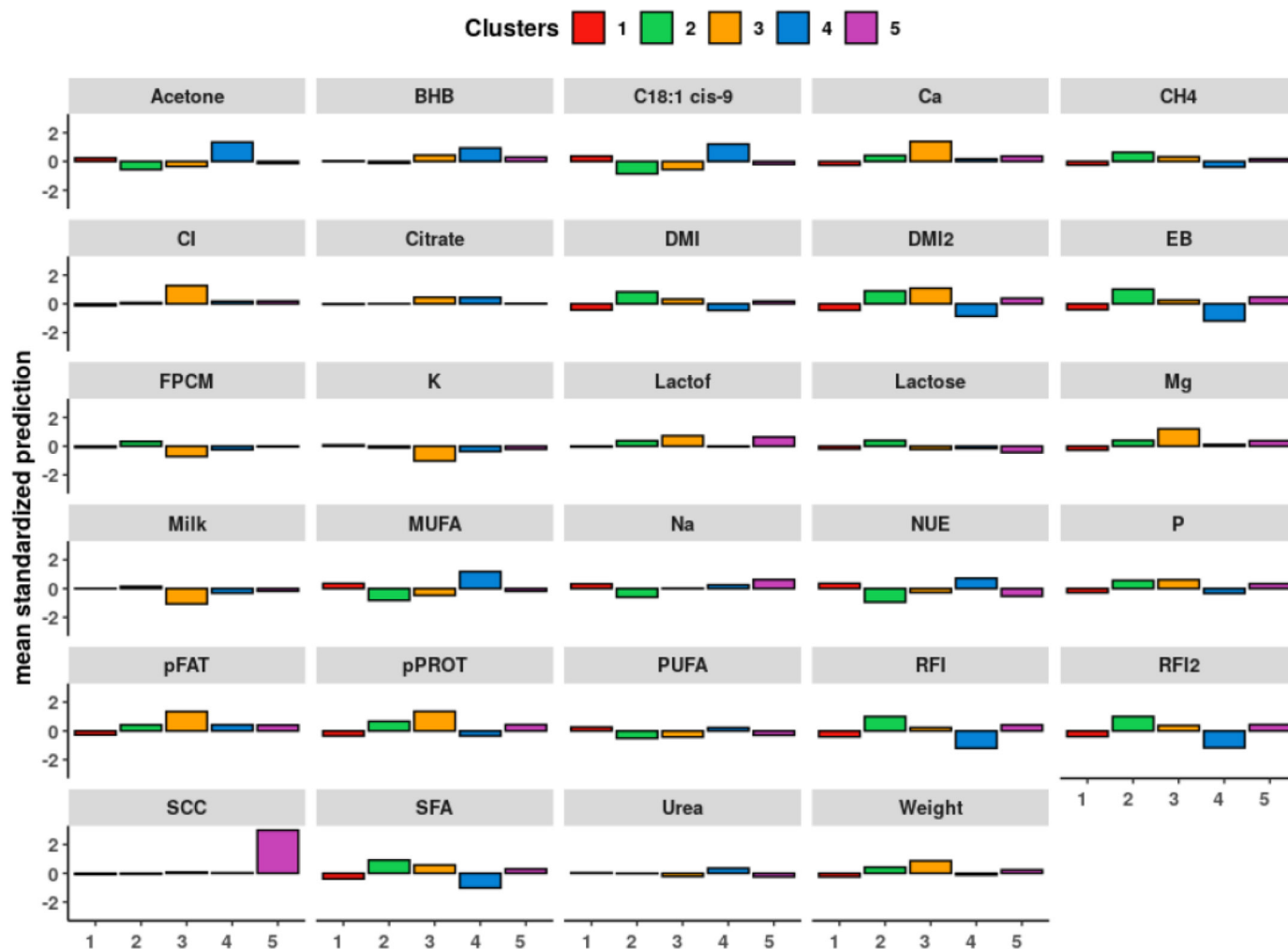
Figure 2. Bar plots of the least squares means of the 29 biomarkers for the 50,000 centroids. C18:1 *cis*-9 = oleic acid; CI = consumption index; EB = energy balance; FPCM = fat- and protein-corrected milk; Lactof = lactoferrin; NUE = nitrogen use efficiency; pPROT = protein content; pFAT = fat content; RFI = residual feed intake (alternate calculation via RFI2); DMI2 = predicted DMI.

smaller cluster. As a result, the training data set was balanced with around the same number of observations for each cluster, creating a model that was not biased toward the prediction of the most represented clusters. The number of latent variables was fixed at 6, as the highest overall cross-validation accuracy was found for this model (0.77). The accuracy decreased to 0.70 when the model was applied to the original centroid data. The balanced accuracy of discriminating cluster 4 and cluster 5 was 0.88 and 0.96, respectively. With use of the PLS-DA model, we were able to extend the prediction of clusters to the entire data set. We observed that cluster 1 was assigned to 21.39% of the records, cluster 2 to 40.89%, cluster 3 to 18.01%, cluster 4 to 16.62%, and cluster 5 to 3.1%. This repartition is more or less the same as the repartition observed using the 50,000

centroids, confirming the model, with a lower percentage of records belonging to cluster 4 and cluster 5. This is expected as those groups should be related to cows with metabolic disorders or mastitis.

Although we had clusters for all records, these categorical variables do not fully represent the reality of a sick animal. Animal health status is a continuous variable from healthy to sick. By observing the PLS-DA latent variables, we can observe this continuum and the overlap between clusters (Figure 3). Therefore, the probabilities of being in cluster 4 or cluster 5 seemed to be more relevant compared with the previous hard clustering, where each observation belongs to only one cluster. However, it is important to verify whether the found clusters were still pertinent using reference biomarker measurements.
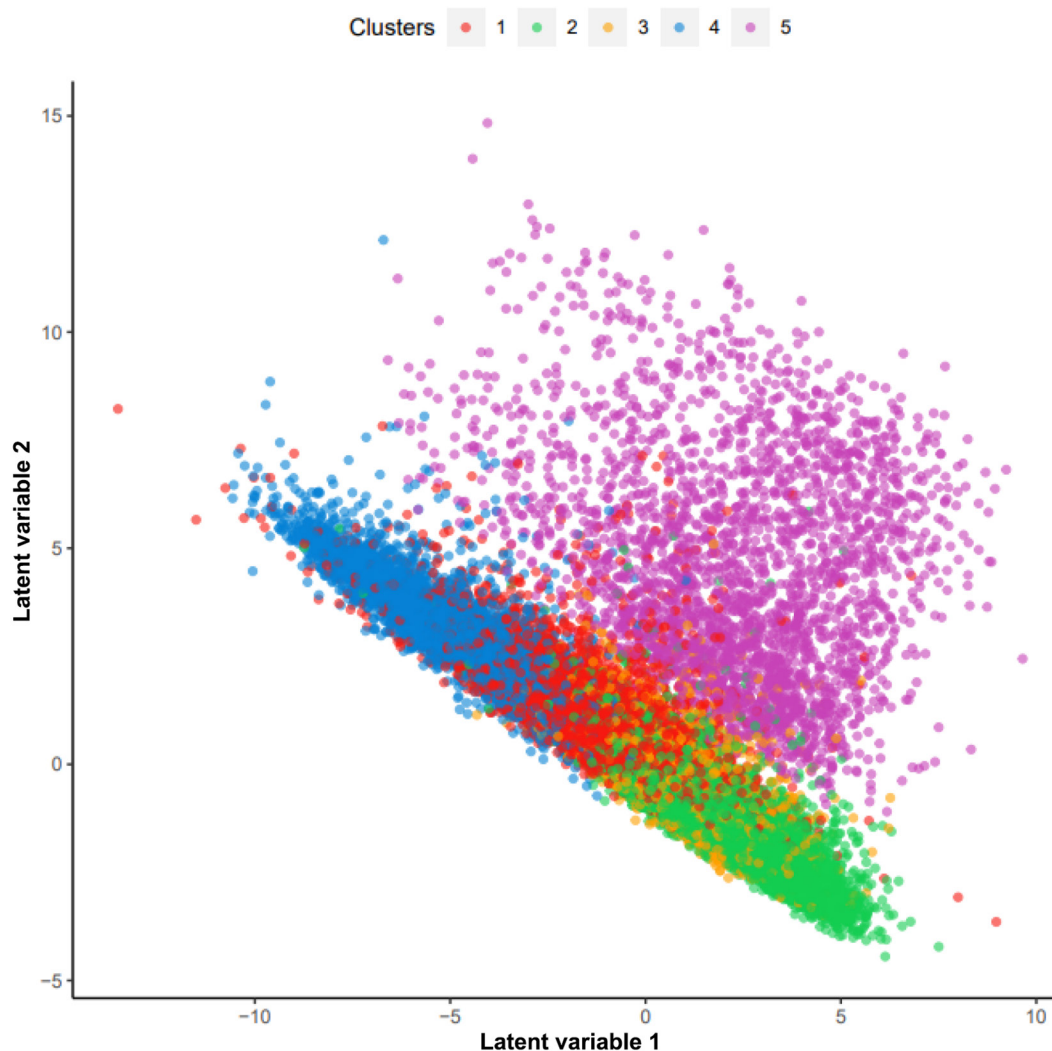
Figure 3. Scatterplot representing the distribution of the 740,454 records on the first and second latent variables from the partial least squares discriminant analysis.

Validating Against Reference Values

Based on the GplusE project, it was possible to create a data set containing reference measurements of biomarkers related to animal health. As all predictors needed to predict the clusters were available for these records, we were able to estimate for each record the probability of being in a cluster from the PLS-DA (Table 5). As the variabilities from the data set composed of primiparous Holsteins were low, especially for the probability related to cluster 5, we assume that this small data set contained very few cases of mastitis. If we consider a proportion of 6% for cluster 5, we would expect 3 observations for this cluster. To solve this issue, we extended the result to subsequent lactations. The extension allowed the consideration of a larger number of observations, where more cows with health

issues could be expected. This example shows the difficulty in finding representative reference data when working with health data.

As all of the PLS-DA probabilities were quantitative, Pearson correlations could be computed between these outputs and the reference contents of biomarkers measured in blood and milk. Table 6 shows these correlations and their significance. The correlations confirmed the previous interpretation of the clusters. Among the significant correlations, the probability associated with cluster 4 was positively correlated with blood and milk BHB, isocitrate, and NEFA and negatively correlated with blood fructosamine, blood glucose, milk free glucose, and milk glucose-6-phosphate. As reported in many studies, these biomarkers reflect EB issues up to ketosis through a lower level of circulating glucose in blood, reflected in the long term through a

Table 5. Descriptive statistics for the probability of being in a cluster for the validation records

Measure	Lactation 1 (n = 50)				All lactations (n = 204)			
	Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD
Cluster 1	0.12	0.31	0.25	0.03	0.03	0.35	0.25	0.04
Cluster 2	0.09	0.27	0.18	0.04	0.01	0.33	0.19	0.05
Cluster 3	0.12	0.19	0.33	0.05	0.03	0.33	0.18	0.04
Cluster 4	0.12	0.55	0.22	0.09	0.04	0.55	0.20	0.08
Cluster 5	0.12	0.27	0.18	0.02	0.12	0.88	0.18	0.09

lower content of fructosamine, and leading to release of NEFA because of body fat mobilization and ultimately production of ketone bodies in the liver (Ingvartsen, 2006; Suthar et al., 2013; Esposito et al., 2014). Larsen (2014) and Larsen and Moyes (2015) also proved that free glucose, glucose-6-phosphate, and isocitrate are related to EB issues or risk of diseases.

For the correlation between the probability of belonging to cluster 5 and the biomarkers, we observed that they differed between the first and the subsequent lactations. We assumed that this was related to the low variability in the first-lactation data set, suggesting that there were no severe mastitis issues in the 50 observations. However, we do not have the on-farm information about mastitis so we cannot validate this assumption. Thus, we also considered the correlations from the other data set. The probability of belonging to cluster 5 was negatively correlated with free glucose in milk but not in blood, and positively correlated with NAGase and LDH. Similarly, Larsen et al. (2010) and Kitchen et al. (1980) showed that LDH and NAGase were relevant indicators of mastitis.

As shown by the validation step based on reference laboratory data, the probabilities were strongly correlated with relevant biomarkers of health issues. Thus, our results highlighted that even using predictions from models with low or fair quality (9 models had an R^2cv < 0.7), trends and patterns can be found in a large data set and validated through external data. As mentioned by McParland and Berry (2016) or Bonfatti et al. (2017), the current results validate the idea that models with low or fair accuracy can be valuable for breeding and monitoring if they can highlight trends or genetic correlation through large-scale data sets. Moreover, the measures could be exploited in a longitudinal approach by computing the variance, the mean value for lactation, or the expected curves. The measures, frequencies, and means of measures could be indicators of the resilience of the animals. However, the method requires more frequently collected spectral data, not a monthly test, to be used as a continuous monitoring tool.

The current approach has the limitation of requiring previous biomarker models to constitute the clusters.

However, the cluster prediction method, being based only on DHI records, opens up many avenues to use already available data. The method offers a way to summarize information from multiple models, to find consistency between all the predictions, and ultimately to detect health patterns in large databases.

CONCLUSIONS

This research aimed to consider the multidimensionality of animal health and its associated traits to detect potential health issues in a routine environment. Instead of using reference laboratory data, the innovative aspect of this study was to make use of a DHI data set containing records related to usual dairy traits and FT-MIR–based predictions that have direct or indirect links with health issues, to discover groups of cows with different data patterns. Even though the method had some computational and validation limitations, this novel approach allowed detection of global patterns from a large data set that is more representative of the population. The results are promising, showing that the approach is worth developing. Two clusters were shown to represent cows with metabolic disorders or that have mastitis issues. From the probability of belonging to 1 of those 2 clusters estimated through a PLS-DA model, we propose a quantitative gradient of sickness related to metabolic disorders or mastitis. These traits are directly usable in many applications; for example, as a tool to monitor the animal health status of every productive cow during routine milk recording or to develop new estimated breeding values due to the large amount of directly available data.

ACKNOWLEDGMENTS

Table 6. Correlations between the probability of being in a cluster and reference measurements of biomarkers related to animal health

Reference analysis[1]	Lactation 1 (n = 50)					All lactations (n = 204)				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Blood										
Cholesterol	0.05	0.33*	−0.17	−0.09	0.02	0.01	0.28***	−0.12	−0.06	−0.07
Fructosamine	0	0.09	0.13	−0.15	0.19	0.03	0.18*	0.01	−0.19**	0.03
BHB	−0.51***	−0.46***	−0.38**	0.70***	−0.4**	−0.34***	−0.31***	−0.1	0.54***	−0.06
NEFA	−0.25	−0.45***	−0.55***	0.67***	−0.36**	−0.14*	−0.39***	−0.13	0.44***	−0.01
P4	0.13	0.21	−0.05	−0.09	−0.11	0.06	0.07	−0.13	0.05	−0.06
Glucose	0.25	0.49***	0.35*	−0.60***	0.42**	0.25***	0.28***	0.08	−0.32***	−0.06
Milk										
Glu6P	0.12	−0.05	0.45***	−0.32*	0.29*	−0.08	−0.08	0.28***	−0.23***	0.16*
GluFree	0.04	0.23	−0.45*	0.16	−0.26	0.31***	0.45***	−0.28***	−0.15*	−0.18**
BHB	−0.6***	−0.34*	−0.17	0.53***	−0.21	−0.34***	−0.24***	−0.04	0.35***	0.04
Isocitrate	−0.46***	−0.61***	0.01	0.5***	−0.16	−0.32***	−0.34***	0.32***	0.38***	−0.1
Urea	−0.15	−0.16	−0.29*	0.32*	−0.15	0.05	0.12	−0.22**	0.03	−0.03
NAGase	−0.09	−0.36**	0.19	0.1	0.02	−0.39***	−0.46***	−0.07	0.06	0.46***
LDH	−0.18	−0.29*	0.34*	−0.07	0.43**	−0.52***	−0.39***	−0.16*	−0.13	0.69***

[1]NEFA = nonesterified fatty acid; P4 = progesterone; Glu6P = glucose-6-phosphate; GluFree = free glucose; NAGase = N-acetyl-β-D-glucosaminidase; LDH = lactate dehydrogenase.

*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

REFERENCES

Atashi, H., M. Salavati, J. De Koster, M. A. Crowe, G. Opsomer, and M. Hostens. 2020. Genome-wide association for metabolic clusters in early-lactation Holstein dairy cows. J. Dairy Sci. 103:6392–6406. https://doi.org/10.3168/jds.2019-17369.

Bansal, B. K., J. Hamann, N. T. Grabowski, and K. B. Singh. 2005. Variation in the composition of selected milk fraction samples from healthy and mastitic quarters, and its significance for mastitis diagnosis. J. Dairy Res. 72:144–152. https://doi.org/10.1017/S0022029905000798.

Bastin, C., L. Théron, A. Lainé, and N. Gengler. 2016. On the role of mid-infrared predicted phenotypes in fertility and health dairy breeding programs. J. Dairy Sci. 99:4080–4094. https://doi.org/10.3168/jds.2015-10087.

Belay, T. K., B. S. Dagnachew, Z. M. Kowalski, and T. Ådnøy. 2017. An attempt at predicting blood β-hydroxybutyrate from Fourier-transform mid-infrared spectra of milk using multivariate mixed models in Polish dairy cattle. J. Dairy Sci. 100:6312–6326. https://doi.org/10.3168/jds.2016-12252.

Blokhuis, H. J., R. B. Jones, R. Geers, M. Miele, and I. Veissier. 2003. Measuring and monitoring animal welfare: Transparency in the food product quality chain. Anim. Welf. 12:445–455.

Bonfatti, V., S.-A. Turner, B. Kuhn-Sherlock, T. D. W. Luke, P. N. Ho, C. V. C. Phyn, and J. E. Pryce. 2019. Prediction of blood β-hydroxybutyrate content and occurrence of hyperketonemia in early-lactation, pasture-grazed dairy cows using milk infrared spectra. J. Dairy Sci. 102:6466–6476. https://doi.org/10.3168/jds.2018-15988.

Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier. 2017. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. J. Dairy Sci. 100:5526–5540. https://doi.org/10.3168/jds.2016-11667.

Bresolin, T., and J. R. R. Dórea. 2020. Infrared spectrometry as a high-throughput phenotyping technology to predict complex traits in livestock systems. Front. Genet. 11:923. https://doi.org/10.3389/fgene.2020.00923.

Christophe, O. S., C. Grelet, C. Bertozzi, D. Veselko, C. Lecomte, P. Höeckels, A. Werner, F.-J. Auer, N. Gengler, F. Dehareng, and H. Soyeurt. 2021. Multiple breeds and countries' predictions of mineral contents in milk from milk mid-infrared spectrometry. Foods 10:2235. https://doi.org/10.3390/foods10092235.

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

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186. https://doi.org/10.3168/jds.2013-6799.

Ebrahimie, E., F. Ebrahimi, M. Ebrahimi, S. Tomlinson, and K. R. Petrovski. 2018. A large-scale study of indicators of sub-clinical mastitis in dairy cattle by attribute weighting analysis of milk composition features: Highlighting the predictive power of lactose and electrical conductivity. J. Dairy Res. 85:193–200. https://doi.org/10.1017/S0022029918000249.

Esposito, G., P. C. Irons, E. C. Webb, and A. Chapwanya. 2014. Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows. Anim. Reprod. Sci. 144:60–71. https://doi.org/10.1016/j.anireprosci.2013.11.007.

Fernando, R. S., S. L. Spahr, and E. H. Jaster. 1985. Comparison of electrical conductivity of milk with other indirect methods for detection of subclinical mastitis. J. Dairy Sci. 68:449–456. https://doi.org/10.3168/jds.S0022-0302(85)80844-4.

Fiore, E., F. Blasi, M. Morgante, L. Cossignani, T. Badon, M. Gianesella, B. Contiero, and M. Berlanda. 2020. Changes of milk fatty acid composition in four lipid classes as biomarkers for the diagnosis of bovine ketosis using bioanalytical thin layer chromatography and gas chromatographic techniques (TLC-GC). J. Pharm. Biomed. Anal. 188:113372. https://doi.org/10.1016/j.jpba.2020.113372.

Fleming, A., F. S. Schenkel, J. Chen, F. Malchiodi, V. Bonfatti, R. A. Ali, B. Mallard, M. Corredig, and F. Miglior. 2017. Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets. J. Dairy Sci. 100:5073–5081. https://doi.org/10.3168/jds.2016-12102.

Foldager, L., C. Gaillard, M. T. Sorensen, T. Larsen, E. Matthews, R. O'Flaherty, F. Carter, M. A. Crowe, C. Grelet, M. Salavati, M. Hostens, K. L. Ingvartsen, and M. A. Krogh. 2020. Predicting physiological imbalance in Holstein dairy cows by three different sets of milk biomarkers. Prev. Vet. Med. 179:105006. https://doi.org/10.1016/j.prevetmed.2020.105006.

Fraser, D. 1995. Science, values and animal welfare: Exploring the 'inextricable connection'. Anim. Welf. 4:103–117.

Gandomi, A., and M. Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manage. 35:137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Gengler, N., H. Soyeurt, F. Dehareng, C. Bastin, F. Colinet, H. Hammami, M.-L. Vanrobays, A. Lainé, S. Vanderick, C. Grelet, A. Vanlierde, E. Froidmont, and P. Dardenne. 2016. Capitalizing on fine milk composition for breeding and management of dairy cows1. J. Dairy Sci. 99:4071–4079. https://doi.org/10.3168/jds.2015-10140.

Grelet, C., C. Bastin, M. Gelé, J.-B. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng. 2016. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. J. Dairy Sci. 99:4816–4825. https://doi.org/10.3168/jds.2015-10477.

Grelet, C., P. Dardenne, H. Soyeurt, J. A. Fernandez, A. Vanlierde, F. Steevens, N. Gengler, and F. Dehareng. 2021. Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. Methods 86:97–111. https://doi.org/10.1016/j.ymeth.2020.07.012.

Grelet, C., J. Fernandez Pierna, H. Soyeurt, F. Dehareng, N. Gengler, and P. Dardenne. 2014. Creation of universal MIR calibration by standardization of milk spectra: example of fatty acids. Pages 108 in Book of Abstracts of the 65th Annual Meeting of the European Federation of Animal Science. Wageningen Academic Publishers.

Grelet, C., E. Froidmont, L. Foldager, M. Salavati, M. Hostens, C. P. Ferris, K. L. Ingvartsen, M. A. Crowe, M. T. Sorensen, J. A. Fernandez Pierna, A. Vanlierde, N. Gengler, and F. Dehareng. 2020. Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation. J. Dairy Sci. 103:4435–4445. https://doi.org/10.3168/jds.2019-17910.

Grelet, C., A. Vanlierde, F. Dehareng, E. Froidmont, and G. Consortium. 2017. Prediction of energy status of dairy cows using MIR milk spectra. Pages 403 in Book of Abstracts of the 68th Annual Meeting of the European Federation of Animal Science, Tallin, Wageningen Academic Publishers.

Grelet, C., A. Vanlierde, M. Hostens, L. Foldager, M. Salavati, K. L. Ingvartsen, M. Crowe, M. T. Sorensen, E. Froidmont, C. P. Ferris, C. Marchitelli, F. Becker, T. Larsen, F. Carter, and F. Dehareng. 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.

Gross, J., H. A. van Dorland, R. M. Bruckmaier, and F. J. Schwarz. 2011. Performance and metabolic profile of dairy cows during a lactational and deliberately induced negative energy balance with subsequent realimentation. J. Dairy Sci. 94:1820–1830. https://doi.org/10.3168/jds.2010-3707.

Ingvartsen, K. L. 2006. Feeding- and management-related diseases in the transition cow: Physiological adaptations around calving and strategies to reduce feeding-related diseases. Anim. Feed Sci. Technol. 126:175–213. https://doi.org/10.1016/j.anifeedsci.2005.08.003.

Kitchen, B. J., G. Middleton, I. G. Durward, R. J. Andrews, and M. C. Salmon. 1980. Mastitis diagnostic tests to estimate mammary gland epithelial cell damage. J. Dairy Sci. 63:978–983. https://doi.org/10.3168/jds.S0022-0302(80)83035-9.

Krogh, M. A., M. Hostens, M. Salavati, C. Grelet, M. T. Sorensen, D. C. Wathes, C. P. Ferris, C. Marchitelli, F. Signorelli, F. Napolitano, F. Becker, T. Larsen, E. Matthews, F. Carter, A. Vanlierde, G. Opsomer, N. Gengler, F. Dehareng, M. A. Crowe, K. L. Ingvartsen, and L. Foldager. 2020. Between- and within-herd variation in blood and milk biomarkers in Holstein cows in early lactation. Animal 14:1067–1075. https://doi.org/10.1017/S1751731119002659.

Larsen, T. 2014. Fluorometric determination of free and total isocitrate in bovine milk. J. Dairy Sci. 97:7498–7504. https://doi.org/10.3168/jds.2014-8018.

Larsen, T., and K. M. Moyes. 2015. Are free glucose and glucose-6-phosphate in milk indicators of specific physiological states in the cow? Animal 9:86–93. https://doi.org/10.1017/S1751731114002043.

Larsen, T., C. M. Røntved, K. L. Ingvartsen, L. Vels, and M. Bjerring. 2010. Enzyme activity and acute phase proteins in milk utilized as indicators of acute clinical *E. coli* LPS-induced mastitis. Animal 4:1672–1679. https://doi.org/10.1017/S1751731110000947.

LeBlanc, S. 2010. Monitoring metabolic health of dairy cattle in the transition period. J. Reprod. Dev. 56(Suppl.):S29–S35. https://doi.org/10.1262/jrd.1056S29.

Luke, T. D. W., S. Rochfort, W. J. Wales, V. Bonfatti, L. Marett, and J. E. Pryce. 2019. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. J. Dairy Sci. 102:1747–1760. https://doi.org/10.3168/jds.2018-15103.

McParland, S., and D. P. Berry. 2016. The potential of Fourier transform infrared spectroscopy of milk samples to predict energy intake and efficiency in dairy cows. J. Dairy Sci. 99:4056–4070. https://doi.org/10.3168/jds.2015-10051.

NRC. 2001. Nutrient Requirements of Dairy Cattle. 7th rev. ed. The National Academies Press.

Pralle, R. S., K. W. Weigel, and H. M. White. 2018. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. J. Dairy Sci. 101:4378–4387. https://doi.org/10.3168/jds.2017-14076.

Smith, S. L., S. J. Denholm, M. P. Coffey, and E. Wall. 2019. Energy profiling of dairy cows from routine milk mid-infrared analysis. J. Dairy Sci. 102:11169–11179. https://doi.org/10.3168/jds.2018-16112.

Soyeurt, H., C. Bastin, F. G. Colinet, V. M.-R. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: Potential indi-

cator of mastitis. Animal 6:1830–1838. https://doi.org/10.1017/S1751731112000791.

Soyeurt, H., C. Grelet, S. McParland, M. Calmels, M. Coffey, A. Tedde, P. Delhez, F. Dehareng, and N. Gengler. 2020. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. J. Dairy Sci. 103:11585–11596. https://doi.org/10.3168/jds.2020-18870.

Suthar, V. S., J. Canelas-Raposo, A. Deniz, and W. Heuwieser. 2013. Prevalence of subclinical ketosis and relationships with postpartum diseases in European dairy cows. J. Dairy Sci. 96:2925–2938. https://doi.org/10.3168/jds.2012-6035.

Tedde, A., C. Grelet, P. N. Ho, J. E. Pryce, D. Hailemariam, Z. Wang, G. Plastow, N. Gengler, Y. Brostaux, E. Froidmont, F. Dehareng, C. Bertozzi, M. A. Crowe, I. Dufrasne, H. Soyeurt, and GplusE Consortium. 2021. Validation of dairy cow bodyweight prediction using traits easily recorded by dairy herd improvement organizations and its potential improvement using feature selection algorithms. Animals (Basel) 11:1288. https://doi.org/10.3390/ani11051288.

Vanlierde, A., F. Dehareng, N. Gengler, E. Froidmont, S. McParland, M. Kreuzer, M. Bell, P. Lund, C. Martin, B. Kuhla, and H. Soyeurt. 2021. Improving robustness and accuracy of predicted daily methane emissions of dairy cows using milk mid-infrared spectra. J. Sci. Food Agric. 101:3394–3403. https://doi.org/10.1002/jsfa.10969.

Vranković, L., J. Aladrović, D. Octenjak, D. Bijelić, L. Cvetnić, and Z. Stojević. 2017. Milk fatty acid composition as an indicator of energy status in Holstein dairy cows. Arch. Tierzucht 60:205–212. https://doi.org/10.5194/aab-60-205-2017.

Wang, C., M.-H. Chen, E. Schifano, J. Wu, and J. Yan. 2016. Statistical methods and computing for big data. Stat. Interface 9:399–414. https://doi.org/10.4310/SII.2016.v9.n4.a1.

Ward, J. H. Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58:236–244. https://doi.org/10.1080/01621459.1963.10500845.

Yan, M.-J., J. Humphreys, and N. M. Holden. 2011. An evaluation of life cycle assessment of European milk production. J. Environ. Manage. 92:372–379. https://doi.org/10.1016/j.jenvman.2010.10.025.