J. Dairy Sci. 100:5073–5081
https://doi.org/10.3168/jds.2016-12102

Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets

A. Fleming,*[1] F. S. Schenkel,* J. Chen,† F. Malchiodi,* V. Bonfatti,‡ R. A. Ali,§ B. Mallard,# M. Corredig,† and F. Miglior*‖

*Centre for Genetic Improvement of Livestock, Department of Animal Biosciences, University of Guelph, Guelph, ON, N1G 2W1, Canada
†Department of Food Science, University of Guelph, Guelph, ON, N1G 2W1, Canada
‡Department of Comparative Biomedicine and Food Science (BCA), University of Padova, Viale dell'Università 16, 35020, Legnaro, Italy
§Department of Mathematics and Statistics, and
#Department of Pathobiology, University of Guelph, Guelph, ON, N1G 2W1, Canada
‖Canadian Dairy Network, Guelph, ON, N1K 1E5, Canada

ABSTRACT

The fatty acid profile of milk is a prevailing issue due to the potential negative or positive effects of different fatty acids to human health and nutrition. Mid-infrared spectroscopy can be used to obtain predictions of otherwise costly fatty acid phenotypes in a widespread and rapid manner. The objective of this study was to evaluate the prediction of fatty acid content for the Canadian dairy cattle population from mid-infrared spectral data and to compare the results produced by altering the partial least squares (PLS) model development set used. The PLS model development sets used to develop the predictions were reference fatty acids expressed as (1) grams per 100 g of fatty acid, (2) grams per 100 g of milk, (3) the natural logarithmic transform of grams per 100 g of milk, and (4) subsets of samples randomly selected by removing excess records around the mean to present a more uniform distribution, repeated 10 times. Gas chromatography measured fatty acid concentration and spectral data for 2,023 milk samples of 373 cows from 4 breeds and 44 herds were used in the model development. The coefficient of determination of cross-validation $\left(R^2_{cv}\right)$ increased when fatty acids were expressed on a per 100 g of milk basis compared with on a per 100 g of fat basis for all examined fatty acids. The logarithmic transformation used to create a more Gaussian distribution in the development set had little effect on the prediction accuracy. The individual fatty acids C12:0, C14:0, C16:0, C18:0, C18:1n-9 *cis*, and saturated, monounsaturated, unsaturated, short-chain, medium-chain, and long-chain fatty acid groups had R^2_{cv} greater than 0.70. When model development was performed with subsets of the original samples, slight increases in R^2_{cv} values were observed for the majority of fatty acids. The difference in R^2_{cv} between the top- and bottom-performing prediction equation across the different subsets for a single predicted fatty acid was on average 0.055 depending on which samples were randomly selected to be used in the PLS model development set. Predictions for fatty acids with high accuracies can be used to monitor fatty acid contents for cows in milk recording programs and possibly for genetic evaluation.

Key words: mid-infrared spectroscopy, fatty acid

INTRODUCTION

Milk and milk products are major contributors of nutrients to the diet of many humans, and consumer awareness of the health effects of milk requires an increased look at its fatty acid profile. Over 400 different fatty acids have been identified in milk fat, although most appear only in trace amounts (Christie, 1995). Bovine milk fat has only about 12 fatty acids present at above 1% concentration (Jensen, 2002). Milk fat is rich in many fatty acids with significance to human health (Haug et al., 2007). An increase in the proportion of fatty acids beneficial to health may therefore be of interest. Beyond the nutritional aspect of fatty acids, the relative concentrations of different fatty acids have implications on the technological properties of milk and milk products (Huppertz and Kelly, 2009). As well, changes in milk fatty acids may be an indicator of cow health and energy status (Stoop et al., 2009). Therefore, strategies for managing or altering the fatty acid content of milk are of interest to the dairy industry, and a practical method for phenotyping this trait is required.

Received October 4, 2016.
Accepted February 21, 2017.
[1]Corresponding author: fleminga@uoguelph.ca

Milk composition testing is routinely performed by certified milk recording laboratories using mid-infrared (**MIR**) spectroscopy for payment, quality control, herd management, and genetic selection purposes. The obtained spectra of milk samples are used to simultaneously provide information on a variety of compositional parameters in a rapid and inexpensive manner. In 2015, approximately 700,000 cows in Canada (72% of all Canadian dairy cows) were enrolled on official or management milk recording programs (Canadian Dairy Information Centre, Agriculture and Agri-Food Canada, 2016). A high volume of data is produced and stored in the form of MIR spectra, which can potentially be exploited to predict additional milk traits. De Marchi et al. (2014) reviewed the expanded use of MIR for milk phenotyping. Mid-infrared spectra-based prediction of fatty acid content in milk has shown some, though varied, success using model development sample sets from other dairy cattle populations (Rutten et al., 2009; De Marchi et al., 2011; Soyeurt et al., 2011; Lopez-Villalobos et al., 2014). Therefore, MIR technology may provide a method of obtaining fatty acid composition phenotypes on large numbers of samples in Canada at a low cost from existing milk recording data.

The samples included in the prediction model development sample set have major implications on the effectiveness of the model in predicting new records. Selecting samples based on spectral characteristics has been a popular method in infrared prediction equation development. This procedure involves recording the spectra, selecting samples likely to provide the best prediction based on the spectral variance, and finally performing reference analysis on these samples only. Mid-infrared predictions of fatty acids performed by Soyeurt et al. (2011) identified samples for their model development set by first examining the spectra, which allowed for analyzing fewer milk samples yet still achieving a high amount of variation. In the present study, milk samples and MIR spectra were obtained during routine milk recording and the high volume of samples, the fast throughput of this process, and the vast geography involved, required samples marked for minor milk constituent analysis to be selected before collection. As well, further work with the recorded milk data is suited to multiple samples per cow. As a result of the sampling technique, the samples included in the model development data set for the fatty acids may have an overabundance of samples with the same composition. Restricting the number of similar samples included in the model development set could better the prediction of some fatty acids.

The objective of this study was to investigate the capability of predicting fatty acid content from MIR spectral data collected during routine Canadian DHI milk recording by altering the model development data set by adjusting the scale and distribution of samples in the model development data set.

MATERIALS AND METHODS

Milk Sampling

Milk samples were collected during routine Canadian DHI milk recording by CanWest DHI (Guelph, ON) from February 2014 to October 2015 and Valacta (Sainte-Anne-de-Bellevue, QC) from February 2014 to May 2015. The 44 participating herds were located in the provinces of Alberta, Ontario, and Quebec with Ayrshire, Brown Swiss, Holstein, or Jersey breeds. From each herd, approximately 10 cows were identified (5 at the beginning and 5 at the middle part of the lactation on the first test) and multiple milk samples through 1 or 2 lactations were collected over the study period (20 mo for Ontario and Alberta herds, and 15 mo for the Quebec herds). Individual cow milk samples (50 mL) were collected by DHI field staff and sent to a DHI laboratory with the preservative Bronopol added as per normal milk recording procedures. At the laboratory, the required quantity of milk needed for DHI milk testing was removed and the remainder was sent to the University of Guelph (Guelph, ON) for additional, fine milk component analysis.

Milk Analysis

Milk MIR spectra were obtained from 1 of 2 MilkoScan FT6000 spectrometers (FOSS, Hillerød, Denmark) at either CanWest DHI or Valacta laboratories following routine milk recording methodology. The specifications of the 2 machines and protocols used by the 2 DHI laboratories were identical. The MIR data for each sample contained 1,060 data points in the infrared range of 5,000 to 900 cm^{-1}. Standardization of the historical spectra between the 2 machines and across time was performed per Bonfatti et al. (2017). In this standardization method, a correction factor was used for each wavenumber to correct for shifts in principal component analysis scores that were observed over time. For the purpose of creating prediction equations, regions 3,105 to 3,444 cm^{-1} and 1,628 to 1,658 cm^{-1} of the MIR spectra were removed due to low signal to noise ratio caused by the high absorption of water. No spectral pretreatments were applied to the spectra.

Milk fat extraction was performed at the University of Guelph (Guelph, ON, Canada) using methods adapted from Christie (1982) and Chouinard et al. (1997). The

milk sample (1 mL) was transferred to an acid-washed (sulfuric acid, A300–212, Fisher Scientific, Fair Lawn, NY) glass test tube with a screw-cap GL25, and 0.15 mL of 25% ammonium hydroxide solution (A669–500, Fisher Scientific) was added. Next, 1 mL of anhydrous ethyl alcohol (Commercial Alcohol, Brampton, ON), 4.5 mL of anhydrous ethyl ether (E138–4, Fisher Scientific), and 4.5 mL of pentane (P399–4, Fisher Scientific) were added and vortexed using a Fisher Vortex Genie 2 (12–812, Fisher Scientific). Two phases were clearly separated after 15 min. The upper layer (0.5 mL) was transferred by glass pipette (13–678–20A, Fisher Scientific) into a 1.5-mL Eppendorf tube (Thermo Fisher Scientific, Waltham, MA, #3451). The solvent was dried by nitrogen gas (grade 4.8, Linde Canada, Guelph, ON, Canada). Then, 1 mL of ethyl ether (E138–4, Fisher Scientific) was added to reconstitute the fat, and 200 μL of methyl acetate (45999–250 mL-F, Sigma-Aldrich, St. Louis, MO) was added to methylate. Next, 10 μL of 25% sodium methoxide solution (156256–25 mL, Sigma-Aldrich) was added, and the reaction occurred for 5 min. Then, 60 μL of saturated oxalic acid (A219–500, Fisher Scientific) in ethyl ether was added to stop the reaction. Eppendorf tubes were centrifuged at 3,340 × *g* at room temperature for 5 min using an Eppendorf centrifuge 5415D (Eppendorf, Hauppauge, NY).

Fatty acid composition was determined using an Agilent Technology (New Castle, DE) model 7890B GC equipped with an automatic on-column injector (Agilent G4513A), and a flame-ionization detector used on a CP-Sil88 fused silica capillary column (CP 7489 100 m × 0.25 mm × 0.2 μm film thickness, Agilent J&W). Column conditions were set up as follows: 40 μL of clear methylated samples were dissolving in 1 mL of hexanes (H302–4, Fisher Scientific). Hydrogen was generated by hydrogen generator (Parker Balston, Lancaster, NY), and was used as the carrier gas at flow rate of 1 mL/min, and nitrogen gas was used as the makeup gas (grade 5.0, Linde Canada, Guelph, ON, Canada). Then, 1 μL of the sample was injected directly cold on-column at an oven temperature of 35°C. After initiation, the column temperature was held at 35°C for 5 min, increased by 14°C/min to 165°C, then increased by 2°C/min to 220°C, and was subsequently held there for 17 min. Identification of fatty acid methyl eater peaks was based on retention time of FAME mix C4-C24 fatty acid standard (Supelco, Bellefonte, PA), and methyl tricosanoate (C17, T9900–500mg, Sigma-Aldrich) was used as an internal standard. Individual fatty acid concentrations were obtained as a percentage of total fatty acids. Along with individual fatty acid content, fatty acids were classified into saturated, unsaturated, monounsaturated, polyunsaturated, short-chain (4 to 10 carbons), medium-chain (11 to 16 carbons), or long-chain (17 to 22 carbons) fatty acid groups.

Data

In the final data set, fatty acid analysis was completed on 2,064 milk samples from 374 cows (average 5.52 samples per cow; range 1 to 12 samples per cow). The large range in the number of samples per cow was due to the frequency of tests for a herd, the condition of the sample at the time of analysis, and the condition of the cow (entering their dry period, being in subnormal health, or leaving the herd). Individual and groups of fatty acids were converted from grams per 100 g of fatty acid to grams per 100 g of milk using the fat content determined by the DHI laboratory during milk recording using MIR spectroscopy. Total milk fat content was only obtainable for 1,952 of the samples and thus fewer samples had fatty acids expressed on a per milk basis. To maintain as much variation as possible, but still remove extreme values, individual fatty acid amounts (g/100 g of fat) more than 5 standard deviations away from the mean were removed. Due to the nature of the fatty acid determination, the entire record was deleted if one value was deemed an outlier. After editing, 2,023 samples with fatty acids were expressed as grams per 100 g of fat and 1,911 samples were expressed as grams per 100 g of milk. The number of cows, herds, and per-cow samples after editing by breed is shown in Table 1.

MIR Prediction Development Sets

Four measures of fatty acid content were evaluated for building fatty acid prediction equations from MIR spectra: fatty acid content measured (1) as grams per 100 g of fatty acid, (2) as grams per 100 g of milk, (3) as the natural logarithm of the grams per 100 g of milk fatty acid content plus a constant of 1 (to include samples with values of 0), and (4) by eliminating samples with fatty acid content coming from overrepresented areas of the fatty acid content [in ln(g/100 g of milk + 1)] distribution. The natural logarithm of fatty acid contents on a milk basis was considered to reduce the skewness of the fatty acid content distribution, which was positively skewed for some fatty acids. The final method of eliminating samples was examined because histograms of the fatty acid concentrations showed that many more samples were closer to the mean concentrations than were in the tails of the distributions (a leptokurtic distribution). For the purpose of building prediction equations, having equal numbers of samples across the full range of fatty acid concentrations would be ideal, but impractical to collect. Prediction equations developed

Table 1. The number of cows, herds, and samples after editing with fatty acid content expressed on a fat (g/100 g of fat) and milk (g/100 g of milk) basis by breed of cow

	In fat (g/100 g of fat)			In milk (g/100 g of milk)		
Breed	Cows	Herds[1]	Records	Cows	Herds[1]	Records
Ayrshire	58	7	431	57	7	412
Brown Swiss	25	3	146	20	3	118
Holstein	230	29	1,206	227	29	1,165
Jersey	60	8	240	58	8	216
Total	373	44	2,023	362	44	1,911

[1]Three herds had multiple breeds.

with sample sets having a Gaussian distribution may cause predictions of future samples to regress toward the mean, a phenomenon known as the Dunne effect (Dunne and Anderson, 1976; Williams, 2001). This effect will be more pronounced in sample sets with very large variance and low correlation between the infrared spectrum and reference values (Williams, 2007). Thus, a more uniform distribution was generated from the original model development samples using a uniform random selection procedure. To create the subset independently for each individual fatty acid or each fatty acid group, the natural log-transformed grams per 100 g of milk fatty acid records were partitioned into bins equal to 1/100th of the range of that particular fatty acid. A maximum of 18 samples per bin were randomly selected from each bin (1% of the total number of samples). So, if a bin had only 18 or fewer samples, then all of the records in that particular bin were included in the subset. The number of samples used in the training set to create the prediction equation using the subsets was therefore far fewer than the other approaches and varied between the different fatty acids depending on the original distribution. This subset selection process was repeated 10 times to create 10 different subsets for each individual fatty acid or group. An examination of the distribution of all created model development sets was performed using the UNIVARIATE procedure in SAS (version 13.1, SAS Institute Inc., Cary, NC).

MIR Prediction Models

All prediction equations were constructed by partial least squares (**PLS**) regression using the PLS procedure of SAS (SAS Institute Inc.). First, all samples with both milk fatty acid data and spectral data available were included in the PLS model development set and the individual and groups of fatty acids were regressed on the spectral data using PLS one at a time. The root mean square error for standardized predictors was examined for each milk sample as a measure of the distance between the data point and the model plane in X-space. Sample spectra with a mean square error greater than 3 standard deviations above the mean value were considered outliers and omitted from the analysis. The PLS procedure with a 10-fold cross-validation was then used on the remaining data to produce the final prediction equation results. In particular, the milk samples were randomly divided into 10 partitions and one at a time a partition was reserved as holdout data for testing and all other samples were used to train the model. This process was repeated until each partition has been predicted in turn, with the validation errors saved each time and then averaged to create the standard error of cross-validation. To account for instability in the prediction model performance due to the random assignment of samples in the cross-validation procedure, the process was repeated 10 times for each prediction and the results were averaged. The same procedure was followed for each of the 10 subsets of fatty acids and the final fitting statistics were averaged across the 10 repeats. The coefficient of determination of cross-validation $\left(\mathbf{R}^2_{\mathbf{cv}}\right)$, which indicates the proportion of the sample variation explained by the regression model, was used to assess how well the prediction fit the data. Additionally, the ratio of performance to deviation (**RPD**), calculated as the ratio of the standard deviation of the PLS model development set to the standard error of cross-validation, was determined as an additional measure of model utility.

RESULTS AND DISCUSSION

Descriptive Statistics

Descriptive statistics of the GC-measured individual and grouped fatty acids expressed on a fat basis (g/100 g of fat) and milk basis (g/100 g of milk) are summarized in Table 2. The fatty acids appearing in the highest concentrations were C16:0, C18:1n-9 *cis*, and C14:0. Sufficient variation was observed in the measured fatty acids to produce prediction equations, which is a signature of sampling from a wide range of management

practices. The coefficient of variation for milk samples in the full data set expressed as grams per 100 g of fat ranged from 6.804 to 220.496%, which is comparable, although on average slightly lower, to that of Soyeurt et al. (2011). Not surprisingly, fatty acids given on a per milk basis showed slightly more variation than those given as per fat due to differences in the fat content of the milk samples. Fatty acids that were on average present in very small quantities had large coefficient of variation values. This trend was particularly true for C22:6n-3, which had the lowest concentration of all measured fatty acids. Approximately 80% of the milk samples had recorded concentrations of zero for this fatty acid, which could in part be due to minimum detection values. Consequently, C22:6n-3 had a highly positively skewed, leptokurtic distribution and the associated mean and standard deviation were greatly affected by the presence of zeros. However, upon omission of almost all recorded zeros, the coefficient of variation decreased to a value similar to that of the other fatty acids.

Prediction Equations

The fitting statistics for all of the prediction models are shown in Table 3. Soyeurt et al. (2011) suggested that prediction equations with $R^2_{cv} > 0.95$ could be used for payment purposes, and equations with an $R^2_{cv} > 0.75$ could be used for animal breeding purposes. However, Cecchinato et al. (2009) showed that despite low prediction R^2 for their MIR predicted milk coagulation properties, the genetic correlation between the measured and predicted values were large and predicted values could be used successfully as indicator traits to genetically improve milk coagulation properties. In the present study, the R^2_{cv} of the predictions of fatty acids ranged from 0.13 to 0.76 for g/100 g of fat and 0.17 to 0.94 for g/100 g of milk. For all examined individual and groups of fatty acids, the R^2_{cv} value increased when fatty acids were expressed per milk compared with per fat. No individual fatty acid achieved a R^2_{cv} of at least 0.70 when expressed as grams per 100 g of fat. The fatty acid groups of saturated, monounsaturated, un-

Table 2. Mean and coefficient of variation of GC-determined fatty acid content on a fat (g/100 g of fat; n = 2,023) and milk (g/100 g of milk; n = 1,911) basis

	In fat (g/100 g of fat)		In milk (g/100 g of milk)	
Fatty acid	Mean	CV	Mean	CV
Individual fatty acid				
C4:0	3.729	18.676	0.157	26.003
C6:0	1.835	35.532	0.078	41.678
C8:0	1.609	37.292	0.067	41.264
C10:0	3.488	27.372	0.147	34.990
C11:0	0.446	48.862	0.019	51.353
C12:0	3.793	29.083	0.161	37.573
C13:0	0.180	79.231	0.008	82.541
C14:0	12.314	17.272	0.519	25.935
C14:1	1.116	32.629	0.047	38.066
C15:0	1.236	24.409	0.052	31.364
C16:0	31.281	13.189	1.324	23.965
C16:1	1.944	23.421	0.082	30.916
C17:0	0.733	21.807	0.031	29.931
C17:1	0.209	54.427	0.009	59.993
C18:0	9.820	24.424	0.411	30.927
C18:1n-9 *trans*	2.229	38.447	0.092	40.148
C18:1n-9 *cis*	18.597	21.701	0.778	27.847
C18:2n-6 *trans*	0.224	55.616	0.009	55.722
C18:2n-6 *cis*	1.962	32.455	0.082	35.503
C18:3n-3	0.727	44.182	0.030	46.910
C18:2 *cis*-9,*cis*-12	0.619	43.448	0.026	45.166
C22:6n-3	0.045	220.486	0.002	222.902
Fatty acid group[1]				
SFA	70.465	6.804	2.974	20.655
MUFA	24.095	17.776	1.001	24.600
PUFA	3.577	23.462	0.149	27.179
UFA	27.672	17.264	1.158	23.917
Short-chain	10.661	16.842	0.449	25.882
Medium-chain	52.311	11.523	2.212	22.905
Long-chain	35.166	18.448	1.471	25.224

[1]Short-chain (4 to 10 carbons), medium-chain (11 to 16 carbons), and long-chain (17 to 22 carbons) fatty acid groups.

Table 3. Fitting statistics of each prediction equation estimating fatty acid concentrations using the model development data sets expressed as g/100 g of fat (F), g/100 g of milk (M), ln(g/100 g of milk + 1) (LN), and the subsets (S)[1]

	N samples				R^2_{cv}				RPD			
Fatty acid	F	M	LN	S	F	M	LN	S	F	M	LN	S
Individual fatty acid												
C4:0	1,984	1,874	1,871	907	0.32	0.66	0.66	**0.73**	1.22	1.71	1.71	**1.94**
C6:0	1,976	1,873	1,873	938	0.18	0.38	0.37	**0.46**	1.11	1.27	1.26	**1.37**
C8:0	1,984	1,863	1,871	977	0.21	0.37	0.39	**0.40**	1.12	1.26	1.29	**1.29**
C10:0	1,985	1,875	1,876	843	0.52	0.66	0.67	**0.75**	1.45	1.72	1.74	**2.00**
C11:0	1,976	1,868	1,868	742	0.13	**0.21**	0.20	0.20	1.07	**1.12**	1.12	1.12
C12:0	1,980	1,872	1,873	836	0.61	0.71	0.72	**0.76**	1.59	1.85	1.89	**2.06**
C13:0	1,976	1,870	1,873	572	0.14	0.19	**0.36**	0.14	1.08	1.11	**1.25**	1.08
C14:0	1,970	1,861	1,859	946	0.60	0.80	0.80	**0.85**	1.59	2.23	2.25	**2.56**
C14:1	1,983	1,877	1,877	902	0.48	0.61	0.61	**0.68**	1.39	1.60	1.59	**1.77**
C15:0	1,982	1,875	1,875	882	0.42	0.61	0.61	**0.67**	1.31	1.61	1.61	**1.74**
C16:0	1,971	1,876	1,876	990	0.64	0.86	0.86	**0.91**	1.67	2.70	2.69	**3.25**
C16:1	1,983	1,875	1,875	836	0.39	0.62	0.63	**0.66**	1.28	1.63	1.65	**1.73**
C17:0	1,975	1,864	1,865	749	0.17	0.53	0.52	**0.58**	1.10	1.46	1.45	**1.55**
C17:1	1,977	1,870	1,870	637	0.14	0.31	0.30	**0.43**	1.08	1.21	1.19	**1.32**
C18:0	1,983	1,865	1,867	900	0.58	0.73	0.73	**0.80**	1.54	1.93	1.94	**2.23**
C18:1n-9 *trans*	1,982	1,875	1,875	903	0.55	0.60	0.61	**0.63**	1.50	1.58	1.59	**1.65**
C18:1n-9 *cis*	1,974	1,873	1,873	746	0.69	0.79	0.78	**0.83**	1.80	2.18	2.11	**2.45**
C18:2n-6 *trans*	1,979	1,872	1,873	456	0.14	**0.17**	0.14	0.13	1.08	**1.10**	1.08	1.07
C18:2n-6 *cis*	1,978	1,874	1,874	780	0.58	0.62	0.62	**0.68**	1.54	1.63	1.63	**1.78**
C18:3n-3	1,981	1,863	1,864	914	0.53	0.58	0.58	**0.61**	1.45	1.54	1.54	**1.60**
C18:2 *cis*-9,*cis*-12	1,981	1,875	1,875	784	0.62	0.65	0.65	**0.72**	1.62	1.70	1.70	**1.91**
C22:6n-3	1,982	1,876	1,876	392	0.16	**0.22**	0.21	0.16	1.09	**1.13**	1.13	1.10
Fatty acid group[2]												
SFA	1,984	1,867	1,874	905	0.76	0.94	0.93	**0.96**	2.05	3.95	3.91	**4.76**
MUFA	1,984	1,874	1,874	837	0.75	0.84	0.83	**0.88**	1.98	2.54	2.44	**2.85**
PUFA	1,973	1,865	1,865	844	0.55	0.66	0.65	**0.72**	1.49	1.71	1.70	**1.91**
UFA	1,984	1,874	1,874	837	0.75	0.84	0.83	**0.87**	1.99	2.54	2.44	**2.83**
Short-chain	1,974	1,870	1,869	870	0.42	0.72	0.73	**0.78**	1.32	1.88	1.94	**2.12**
Medium-chain	1,973	1,876	1,875	1,005	0.72	0.90	0.89	**0.92**	1.87	3.09	3.02	**3.54**
Long-chain	1,975	1,873	1,873	868	0.72	0.83	0.81	**0.85**	1.89	2.43	2.32	**2.62**

[1]Bold face represents the model with the highest value. R^2_{cv} = coefficient of determination of cross validation; RPD = ratio of performance deviation.

[2]Short-chain (4 to 10 carbons), medium-chain (11 to 16 carbons), and long-chain (17 to 22 carbons) fatty acid groups.

saturated, medium-chain, and long-chain had R^2_{cv} values greater than 0.70. When prediction models were created for fatty acids on a per milk basis, 5 individual fatty acids had R^2_{cv} values over 0.70. These findings are in line with Soyeurt et al. (2006) and Rutten et al. (2009), who also observed higher accuracies when fatty acids were expressed on a per milk basis versus per fat basis. This is explained partly by the fact that milk samples contain different amounts of total fat and samples with the same relative concentrations of fatty acids can contain very different total quantities of the fatty acids. The MIR spectrum correlates to a greater extent with the total amount of a fatty acid than the proportion of fatty acids.

These results can also be observed from the related RPD values. For RPD a higher value is desired and models with an RPD greater than 2 are said to produce good predictions (De Marchi et al., 2011). Manley (2014) reported that the RPD attempts to scale the error in prediction with the standard deviation of the property. They state RPD values greater than 3 are useful for screening, values greater than 5 can be used for quality control, and values greater than 8 can be used for any application. In the present study, only saturated and medium-chain fatty acids had RPD values greater than 3 on a per milk basis and the majority do not surpass RPD values of 2. Therefore, the application of many of the presented predictions may be limited.

In most cases, the individual or groups of fatty acids examined in the present study that appeared in greater concentrations had the highest R^2_{cv}. The 11 individual or grouped fatty acids that were most prevalent in the milk samples, were the only ones for which $R^2_{cv} > 0.70$ when expressed as grams per 100 g of milk. As well, the fatty acids appearing in negligible amounts did not predict well enough to be useful. The relationship between the fatty acid concentration and predictive model performance has also been identified and discussed by Soyeurt et al. (2006) and De Marchi et al. (2011). Rutten et al. (2009) modeled the relationship

between fatty acid concentration (g/dL of milk) and prediction R^2 and reported an R^2 value of 0.64.

Compared with De Marchi et al. (2011), the present study had greater prediction accuracies for most of the fatty acids examined by both, but the former results were based on a smaller number of milk samples (n = 267) and only Brown Swiss cows. The one exception was C8:0, which predicted poorly in our samples. Soyeurt et al. (2011) used a diverse fatty acid data set of 517 samples, and developed prediction equations with larger R^2_{cv} than reported here for examined fatty acids, with models performing marginally or considerably better. With a larger number of samples (n = 3,622) studied by Rutten et al. (2009), validation R^2 reported for fatty acids predictions were also greater than observed presently. Most notably the short-chain fatty acids in our data were predicted unsatisfactorily, whereas Rutten et al. (2009) observed validation R^2 values greater than 0.90 for all C4:0, C6:0, and C8:0. Ferrand et al. (2011) also showed that these short-chain fatty acids could predict well.

The better results in other studies could be a result of differences in the variability in the model development data set, and the statistical procedures used (De Marchi et al., 2014). The procedures used for GC measurement of fatty acids could affect the end accuracy of the predictions. Also, importantly, the sample set used to develop the prediction equations needs to incorporate all of the variation expected to be in the population to be predicted. Wojciechowski and Barbano (2016) attained sample variation for the development of PLS models for fatty acid chain length and unsaturation by including bulk tank milk from individual herds in different regions, individual cow milk samples at different lactation stages, and modified milk calibration samples. They achieved R^2_{cv} of 0.78 and 0.90 for average chain length and unsaturation, respectively. The methodologies for developing the prediction equations and the different pretreatments that other studies have tried on the received spectra could also create differences. Improved accuracies of prediction equations have been achieved by other studies by using first-derivative preprocessing of the spectra or wavelength selection before PLS regression (Soyeurt et al., 2011; Ferrand-Calmels et al., 2014).

Logarithmic Transformation Development Set

Partial least squares regression methods perform best with data that are fairly symmetrically distributed (Wold et al., 2001). The distributions of the measured fatty acids used in the model development were generally not symmetrical. The skewness ranged between −0.498 and 2.597 with an average skewness of 0.688 for fatty acids in grams per 100 g of fat. In most cases, the skewness of the distribution was lower when fatty acids were expressed as grams per 100 g of fat compared with grams per 100 g of milk (range 0.493 to 2.647, average 1.059). Just over half of the examined individual and groups of fatty acids had skewness greater than 1 when expressed per milk. After a natural logarithmic transformation was performed to produce the PLS model development data set, the skewness ranged from 0.035 to 2.628 and averaged 0.827, with 8 fatty acids still having skewness greater than 1. Some improvement occurred in the amount of skewness, although most of the fatty acids still had nonsymmetric distributions.

However, the performance of the prediction models based on the log-transformed development sets did not greatly affect the R^2_{cv} and RPD values (Table 3), with the exception of that for C13:0. The log-transformation on C13:0 improved the model R^2_{cv}, although the value still remained very low. The fatty acids with the most skewed distributions are those detected in small quantities in milk and did not have adequate prediction to start, and thus, no improvements were noted.

Sample Subset Model Development Set

For most fatty acids, the created subsets succeeded in creating more uniform distributions suitable for prediction equation development. The skewness of the subsets ranged from 0.029 to 1.674 and averaged 0.495. The distributions of those fatty acids appearing mostly in small quantities were still positively skewed due to a lack of available samples in the tail of the distribution. The R^2_{cv} of the predictive models improved when using the subset data for all but 4 fatty acids (Table 3). These fatty acids were C11:0, C13:0, C18:2n-6 *trans*, and C22:6n-3, all of which had R^2_{cv} values less than 0.5 and thus seemed unsatisfactory for prediction.

The subset models were repeated 10 times each with a different randomly selected subset. The R^2_{cv} achieved by the 10 repeats were not identical and the differences between them varied depending on the component. The individual fatty acid C22:6n-3 exhibited the most dramatic differences in R^2_{cv} between the 10 repeats with a range of 0.15. This is likely due to the much smaller sample size of 392 milk samples used, a resultant of the very large number of samples having no quantifiable concentration. However, the performance of the prediction of this fatty acid is far below the level of being useful for all tested development sets and this large R^2_{cv} range is not indicative of problems for other fatty acids. Rutten et al. (2010) examined the relationship between the number of samples used in the model development

and the validation R^2 in MIR predicted fatty acids. When using a small number of samples (n = 100) they observed a large range in R^2 values from 0.05 to 0.30 for C16:0, but as the number of samples increased, the magnitude of the observed range decreased. The average range in R^2_{cv} values observed between the different subsets in the present study was 0.057. In general, it is expected that increasing the number of samples within a development set will produce better predictions and more robust models. Rutten et al. (2010) also observed an improvement in their R^2 with increasing sample numbers. The current study largely saw an increase in R^2_{cv} when the number of samples was decreased. The reduced sample numbers in the model development sets were still sufficiently large for the most part, but these results illustrate the importance of which samples are included in the development set. Note that the number of unique or influential samples could be more important than the total number of samples. However, external validation will be required to satisfactorily identify the precision of the present models and how well they perform on another population of milk samples.

An alternate method for selecting subsets that take into consideration the sample spectra itself should be examined in the future. Assuming that samples with like MIR spectra are compositionally similar, perhaps samples with near identical spectra can be removed from the development set, as they are not contributing new information. This selection process would lessen the Gaussian distribution of model development samples and largely eliminate the Dunne effect, which is why this method is attractive for selecting samples for analytic analysis. However, due to the complex nature of the composition of milk, it may still be challenging to uncover samples differing in a minor milk component of interest if it does not dominate the spectra. Additionally, specific milk components may have different ideal model development sets. By randomly selecting samples out of a group with near identical quantities of one fatty acid, inadvertently, samples with similar composition for another, possibly correlated milk trait such as total fat content may be selected together. This could cause the predictive models to inappropriately put strength on the regions of the spectrum relating to the other component and incorporate the correlation in the model. When measured milk components are readily available, along with spectral data, coupling sample selection strategies involving both sample composition and spectral information could be an improvement upon randomly selecting samples with similar composition to produce the more uniform distribution. Such selection strategies may aid in ensuring variability in regard to other milk components within a group of samples with the same quantity of the fatty acid of interest. As a substitute for spectra, selecting samples based on the trait of interest while also considering other known, measured milk component traits influencing the spectrum could also be investigated. In cases where the prediction model development set is skewed, such as the case with the present fatty acid data, it may be worthwhile to look at alternatives to PLS regression, such as partial quantile regression.

CONCLUSIONS

We examined the use of MIR spectra to predict fatty acid content in bovine milk for the Canadian dairy population. The accuracy of the predictions depended on the fatty acid examined and the PLS model development set used to create the equation. The greatest R^2_{cv} was achieved for fatty acids with high concentrations in milk and when they were expressed on a per milk weight basis. Excluding excess samples from overabundant regions of the distribution made further improvements to the equations and can be further investigated. The predictions for some of the fatty acids are sufficient for monitoring changes in fatty acid profiles and for use in animal breeding programs for potential genetic changes. Predicted fatty acids from equations with lower R^2_{cv} may still be useful as indicators for actual fatty acid contents. Future research will further examine the ideal model development set for different fatty acids and spectral pretreatment procedures to improve prediction equations as well as their utility in genetic improvement programs.

ACKNOWLEDGMENTS

All dairy producers participating in this project are gratefully acknowledged. We warmly acknowledge Ian Rumbles (now at Dairy Record Processing Center, Raleigh, NC) and all the CanWest DHI team (Guelph, ON) and Daniel Lefebvre and the full team at Valacta (Ste-Anne-de-Bellevue, QC, Canada) for kindly organizing the selection of herds and collecting samples for the project. This study was partly funded by the DairyGen council of Canadian Dairy Network (Guelph, ON) and the Natural Sciences and Engineering Research Council of Canada (Ottawa, ON, Canada). This project is also supported by a contribution from the Dairy Research Cluster Initiative (Dairy Farmers of Canada, Agriculture and Agri-Food Canada, the Canadian Dairy Network and the Canadian Dairy Commission). We warmly acknowledge FOSS (Hillerød, Denmark) for partial funding and technical support.

REFERENCES

Bonfatti, V., A. Fleming, A. Koeck, and F. Miglior. 2017. Standardization of milk infrared spectra for the retroactive application of calibration models. J. Dairy Sci. 100:2032–2041.

Canadian Dairy Information Centre, Agriculture and Agri-Food Canada. 2016. Enrollments on milk recording. Accessed May 27, 2016. http://www.dairyinfo.gc.ca/index_e.php?s1=dff-fcil&s2=mrr-pcle&s3=emr-ipcl.

Cecchinato, A., M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J. Dairy Sci. 92:5304–5313.

Chouinard, P. Y., V. Girard, and G. J. Brisson. 1997. Performance and profiles of milk fatty acids of cows fed full fat, heat-treated soybeans using various processing methods. J. Dairy Sci. 80:334–342.

Christie, W. W. 1982. A simple procedure for rapid transmethylation of glycerolipids and cholesterol esters. J. Lipid Res. 23:1072–1075.

Christie, W. W. 1995. Composition and structure of milk lipids. Pages 1–36 in Advanced Dairy Chemistry, Volume 2: Lipids, 2nd ed. P. F. Fox, ed. Chapman & Hall, London, UK.

De Marchi, M., M. Penasa, A. Cecchinato, M. Mele, P. Secchiari, and G. Bittante. 2011. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal 5:1653–1658.

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.

Dunne, W., and J. A. Anderson. 1976. A system for segregating Canadian wheat into subgrades of guaranteed protein content. Can. J. Plant Sci. 56:433–450.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet, F. Faucon, H. Larroque, O. Leray, J. M. Trommenschlager, and M. Brochard. 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemometr. Intell. Lab. 106:183–189.

Ferrand-Calmels, M., I. Palhière, M. Brochard, O. Leray, J. M. Astruc, M. R. Aurel, S. Barbey, F. Bouvier, P. Brunschwig, H. Caillat, M. Douguet, F. Faucon-Lahalle, M. Gelé, G. Thomas, J. M. Trommenschlager, and H. Larroque. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35.

Haug, A., A. T. Høstmark, and O. M. Harstad. 2007. Bovine milk in human nutrition–A review. Lipids Health Dis. 6:25.

Huppertz, T., and A. L. Kelly. 2009. Properties and constituents of cow′s milk. Pages 23–43 in Milk Processing and Quality Management. A. Y. Tamime, ed. Blackwell Publishing, New Delhi, India.

Jensen, R. G. 2002. Invited review: The composition of bovine milk lipids: January 1995 to December 2000. J. Dairy Sci. 85:295–350.

Lopez-Villalobos, N., R. J. Spelman, J. Melis, S. R. Davis, S. D. Berr, K. Lehnert, S. E. Holroyd, A. K. MacGibbon, and R. G. Snell. 2014. Estimation of genetic and crossbreeding parameters of fatty acid concentrations in milk fat predicted by mid-infrared spectroscopy in New Zealand dairy cattle. J. Dairy Res. 81:340–349.

Manley, M. 2014. Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. Chem. Soc. Rev. 43:8200–8214.

Rutten, M. J. M., H. Bovenhuis, K. A. Hettinga, H. J. F. van Valenberg, and J. A. M. van Arendonk. 2009. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 92:6202–6209.

Rutten, M. J. M., H. Bovenhuis, and J. A. M. van Arendonk. 2010. The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. J. Dairy Sci. 93:4872–4882.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667.

Stoop, W. M., H. Bovenhuis, J. M. L. Heck, and J. A. M. van Arendonk. 2009. Effect of lactation stage and energy status on milk fat composition of Holstein-Friesian cows. J. Dairy Sci. 92:1469–1478.

Williams, P. 2007. Sampling, sample preparation, and sample selection. Pages 267–294 in Handbook of near-infrared analysis. 3rd ed. D. A. Burns and E. W. Ciurczak, ed. CRC Press, Boca Raton, FL.

Williams, P. C. 2001. Implementation of near-infrared technology. Pages 145–169 in Near-Infrared Technology in the Agricultural and Food Industries. 2nd ed. P. C. Williams and K. Norris, ed. American Association of Cereal Chemists, St. Paul, MN.

Wojciechowski, K. L., and D. M. Barbano. 2016. Prediction of fatty acid chain length and unsaturation of milk fat by mid-infrared milk analysis. J. Dairy Sci. 99:8561–8570.

Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemometr. Intell. Lab. 58:109–120.