Comparison of different methods to validate a dataset with producer-recorded health events

F. Miglior[1,2]*, A. Koeck*[3]*, D. F. Kelton*[4] *and F. S. Schenkel*[3]

[1]*Guelph Food Research Centre, Agriculture and Agri-Food Canada,*
Guelph, Ontario, Canada, N1G 5C9
[2]*Canadian Dairy Network, Guelph, Ontario, Canada, N1K 1E5*
[3]*Centre for Genetic Improvement of Livestock, University of Guelph,*
Guelph, Ontario, Canada, N1G 2W1
[4]*Department of Population Medicine, Ontario Veterinary College,*
University of Guelph, Guelph, Ontario, Canada, N1G 2W1

Abstract

This study is part of a larger project whose overall objective is to develop genetic evaluations for resistance to mastitis and other diseases in Canada. Health data recorded by producers were available from the National Dairy Cattle Health System. Eight diseases are recorded by producers on a voluntary basis: mastitis, displaced abomasum, ketosis, milk fever, retained placenta, metritis, cystic ovaries and lameness. Data validation is an important part of analysis of producer-recorded health data. The objective of this study was to investigate the impact of 5 data validation methods on genetic evaluations for mastitis resistance in first lactation Canadian Holstein cows. As expected the number of usable herds and the number of records were increased with a less stringent data validation, whereas mastitis frequency was decreased. For genetic analyses, univariate and bivariate linear sire models were fitted. Heritability of mastitis was decreased with a less stringent data validation with estimates ranging from 0.013 to 0.026. Lactation mean somatic cell score was highly correlated with mastitis (0.65-0.69) independently of the applied data validation method for mastitis. Pearson correlations between sire breeding values for mastitis resistance based on the different data validation methods were all higher 0.95. Overall, the present study showed that genetic evaluations for mastitis resistance stay similar independently of the applied data validation method. However, there is evidence for some underreporting of mastitis cases in the Canadian health recording system. Therefore, future work is necessary to increase data quality in the Canadian health recording system.

Keywords: mastitis, data validation, Canadian Holsteins.

Introduction

In Canada, a national dairy cattle health and disease data management system started in 2007. The main objectives of this initiative are to provide information to dairy producers and their veterinarians for herd management and to establish a national genetic evaluation system for genetic selection for disease resistance. Eight diseases that are known to affect herd profitability are recorded by producers on a voluntary basis: mastitis, displaced abomasum, ketosis, milk fever, retained placenta, metritis, cystic ovaries and lameness. The feasibility of using producer recorded health data for genetic evaluations for disease resistance in Canada has been previously shown by Neuenschwander *et al.* (2012) and Koeck *et al.* (2012).

In order to obtain reliable and accurate genetic evaluations, recording of disease cases should be as complete as possible on all participating farms. However, data quality can vary among farms and even for a given farm over time.

In the Scandinavian countries and Austria disease recording systems are implemented on a large scale and genetic evaluations for health traits are carried out routinely. Norway has a mandatory disease recording system for dairy cattle and, therefore, it is assumed that all herds report complete health data. Previously only herds with at least 1 recorded mastitis case per herd and year were considered for routine genetic evaluation in Norway. However, due to small herd size and declining disease frequencies in recent years, this data edit is not applied anymore (B. Heringstad, personal communication). Also, in Sweden, disease recording is mandatory and records from all cows are included in routine genetic evaluation (NAV, 2012). In contrast, Finland and Denmark include only records from herds that participate actively in health recording. In Finland only herds with at least 1 veterinary diagnose per herd and year are included (Y. Pösö, personal communication). In Denmark the herd is considered to be participating if the number of treatments is greater than or equal to 0.3 per calving in the period from calving to 4 or 9 months after calving. In the 9-month period, it is not allowed to be a 3 month period after birth, where there is no reported disease diagnosis in the herd. It is also a requirement that there is at least 7 and 10 calvings in the following 4 - and 9-month period, respectively (U. S. Nielsen, personal communication). In the routine genetic evaluation in Austria, only farms with a minimum average of 0.1 first diagnoses per cow and year are considered. Besides, continuous submission of health data by veterinarians or performance recording technicians is checked (C. Egger-Danner, personal communication).

In the present study the impact of 5 data validation methods on genetic evaluations for mastitis resistance in Canadian Holsteins is presented. The results should lead to a better understanding of data quality in the Canadian health recording system.

Material and methods

Health database

Health data from April 2007 to January 2013 were obtained from the Canadian Dairy Network (Guelph, Ontario). Summary of current data in the database is given in Table 1. The database consisted of 633,876 disease cases from 6,327 herds. Recording of mastitis was done in the majority of herds (88%), followed by displaced abomasum (66%) and retained placenta (62%).

The number of reported disease cases per year and month has shown a continuous increase from 2007 to 2010 and stabilized in the year 2011 (Figure 1). In contrast, the total number of herds recording health data remained almost unchanged in the last 5 years (Figure 2). In 2012, about 4,000 herds recorded health data, which accounts for 42% of all herds under milk recording.

Table 1. Summary statistics of the health traits database.

Health category	Health event	% of disease cases	% of herds
Mastitis	Mastitis	42.1	88
Displaced abomasum	Displaced abomasum	4.7	66
Ketosis	Ketosis	3.0	39
Milk fever	Milk fever	3.6	51
Retained placenta	Retained placenta	8.4	62
Metritis	Acute metritis	5.9	41
	Purulent discharge	3.9	22
	Endometritis	1.4	11
	Chronic metritis	2.4	23
Cystic ovaries	Cystic ovaries	12.1	49
Lameness	Lameness	12.1	57
	Foot rot, laminitis, sole ulcer and other claw disorders	0.4	7

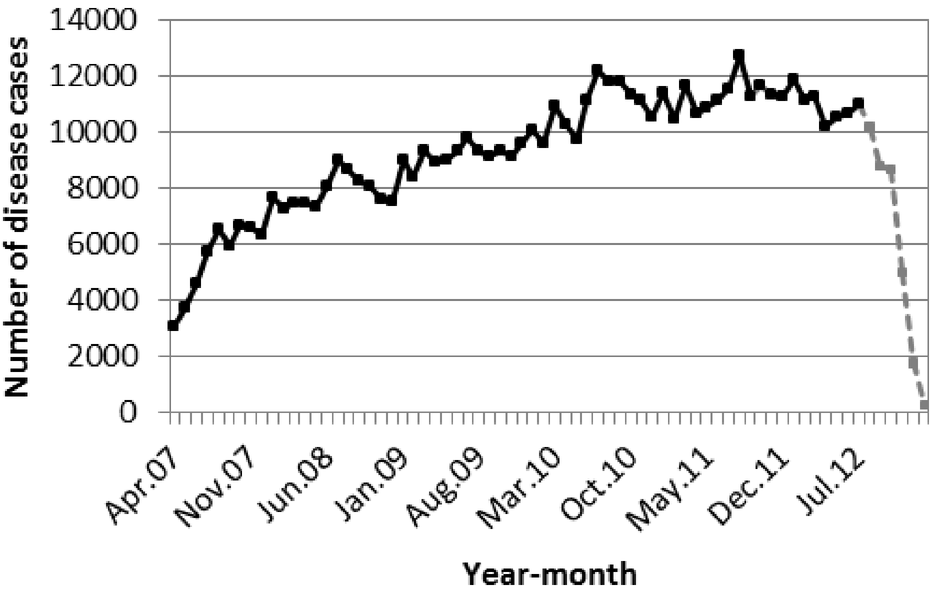

Figure 1. Number of reported disease cases per year and month (dashed line represents delay in data delivery).

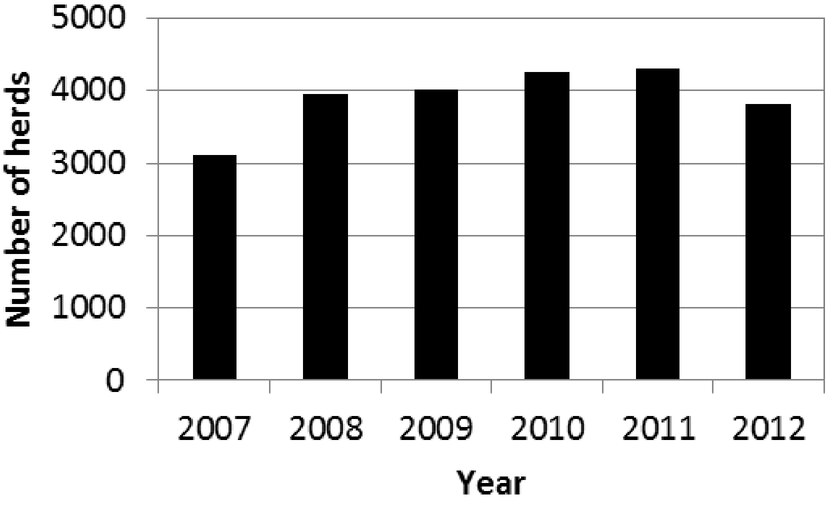

Figure 2. Number of herds recording disease cases per year.

Data validation and editing

For investigating the impact of data validation on number of records, mastitis frequency and genetic parameters 5 different data validation methods were analyzed. In the first 4 scenarios only herds with at least one recorded mastitis case were considered. The first recorded mastitis case was defined as the beginning of the data recording period. Additionally minimum mastitis frequencies per herd and year were applied for data validation methods A, B and C:

- Minimum mastitis frequency of 5% per herd and year (**Method A**)
- Minimum mastitis frequency of 3% per herd and year (**Method B**)
- Minimum mastitis frequency of 1% per herd and year (**Method C**)
- No minimum mastitis frequency per herd and year (**Method D**)

In the last scenario all herds with at least one recorded disease case (any disease) were considered. The first recorded disease case was defined as the beginning of the data recording period. To ensure a continuous data recording a minimum disease frequency of 5% per herd and year was also considered (**Method E**).

Holstein is the most common dairy cattle breed in Canada (constituting up to 90% of the dairy cows) and, therefore, almost all health records were from Holstein cows. For this reason, analyses were carried out for this breed only. For the analyses, only records from first parity cows were considered.

Trait definition

In Canada clinical mastitis cases are recorded by producers. Mastitis is defined as visually abnormal milk (e.g. clots, flakes, or watery) from one or more quarters, that may also include inflammation of the udder (e.g. heat, swelling, or discoloration) and systemic illness of the cow (Kelton *et al.*, 1998).

For analyses, mastitis was defined as a binary trait (0 = no mastitis, 1 = mastitis) based on whether or not the cow had at least one mastitis case in the period from calving to 305 d after calving. Test-day records between 5 and 305 DIM were considered for SCC. Test-day SCC was $\log_2$ transformed to SCS ($\mathrm{SCS} = \log_2(\mathrm{SCC}/100{,}000) + 3$) and averaged over lactation (**LSCS**).

Pedigree

The sire pedigree file was generated by tracing the pedigrees of sires and maternal grandsires back as far as possible.

Model

Linear sire models were fitted using the AI-REML procedure in the DMU package (Madsen and Jensen, 2008). Initially univariate models were run for mastitis. Subsequently bivariate models were run between mastitis and LSCS. In matrix notation, the model was:

$$y = X\beta + Z_h h + Z_s s + e$$

where **y** is a vector of observations for mastitis and LSCS; β is a vector of systematic effects, including fixed effects of age at calving and year-season of calving; h is a vector of random herd-year of calving effects; s is a vector of random additive genetic sire effects; e is a vector of random residuals; and X, Z_h and Z_s, are the corresponding incidence matrices. Random effects were assumed to be normally distributed with zero means, and Var(s)=A σ_s^2, Var(h)=I σ_h^2, Var(e)=I σ_e^2, where σ_s^2, σ_h^2 and are the additive genetic sire, herd-year, and residual variances, respectively, I is an identity matrix, and A is the additive genetic relationship matrix.

Age at first calving had 16 classes, in which <22 and >35 months were the first and last class, respectively, and other classes were single months. Four seasons of calving were defined from January to March, April to June, July to September and October to December.

Results and discussion

Descriptive statistics

Table 2 gives an overview of the data validation methods. As expected the number of usable herds and the number of records were increased with a less stringent data validation, whereas mastitis frequency was decreased. Mastitis frequency was 12.7% based on the most stringent data validation method (A) and in agreement with previous studies. In a literature review, Kelton *et al.* (1998) found a mastitis frequency of 14.2% across studies. A higher mastitis frequency of 20% was obtained by Zwald *et al.* (2004) in US Holstein cows. In a more recent study, Mrode *et al.* (2012) reported a mastitis frequency of 13.5% in first lactation UK Holstein cows.

Mastitis frequencies based on data validation methods B, C, D and E were lower, which possibly indicates underreporting of mastitis cases in the Canadian health recording system.

Table 2. Impact of data validation on number of usable herd, records and mastitis frequency in first lactation Holstein cows.

	Data validation method				
	A[1]	B[2]	C[3]	D[4]	E[5]
Total herds, n	5,076	5,076	5,076	5,076	5,728
Usable herds, n (%)	2,995 (59)	3,342 (66)	3,697 (73)	5,076 (100)	4,110 (72)
Records, n	129,091	151,575	181,405	295,673	217,196
Sires with • 1 daughter, n	5,042	5,426	5,924	7,406	6,301
Sires with • 30 daughters, n	509	634	818	1,626	1,142
Mastitis frequency, %	12.7	11.4	10.0	6.5	8.6
LSCS	2.2	2.2	2.2	2.2	2.2

[1] A: At least one recorded mastitis case and a minimum mastitis frequency of 5% per herd and year.

[2] B: At least one recorded mastitis case and a minimum mastitis frequency of 3% per herd and year.

[3] C: At least one recorded mastitis case and a minimum mastitis frequency of 1% per herd and year.

[4] D: At least one recorded mastitis case.

[5] E: At least one recorded disease case (any disease) and a minimum disease frequency of 5% per herd and year.

Genetic parameters

Heritabilities and genetic correlations for mastitis and LSCS from bivariate analyses are given in Table 3. Generally, heritability estimates of mastitis (0.013 to 0.026) were in the range of previously published studies (e.g. Carlén *et al*., 2004; Mrode *et al*., 2012). Heritability of mastitis was decreased with a less stringent data validation method, whereas heritability of LSCS was nearly constant in the different analyses. The decreasing heritability estimate of mastitis can be partly explained by the decreasing mastitis frequency, as heritability estimates are frequency-dependent when applying linear models to binary data. However, heritability of mastitis was also decreased when transformed to the underlying scale using the classical formula of Dempster and Lerner (1950).

Lactation mean somatic cell score was highly correlated with mastitis (0.65-0.69) independently of the applied data validation method for mastitis.

Table 3. Impact of data validation on genetic parameters from bivariate linear sire models.

	Data validation method				
	A	B	C	D	E
Heritability, Mastitis	0.026	0.022	0.020	0.013	0.017
Heritability, Mastitis_DL[1]	0.067	0.060	0.059	0.049	0.054
Heritability, LSCS	0.123	0.125	0.123	0.122	0.124
Genetic correlation	0.69	0.66	0.65	0.66	0.68

[1] Heritability estimates were transformed to the underlying scale using the formula of Dempster and Lerner (1950).

Estimated breeding values

Pearson correlations between sire breeding values for mastitis resistance based on the different data validation methods are shown in Table 4. The correlations were all higher 0.95 showing that genetic evaluations stay similar across the data validation methods. Although in method E data validation was done across all diseases, genetic evaluation based on this method was highly correlated with the other evaluations. This can be expected, as mastitis is recorded in almost all participating herds.

Table 5 presents the number of top 100 bulls based on data validation method A that are in common with the other methods. Selecting the top 100 bulls based on any data validation method would not have major consequences on selection decisions. As expected the number of top 100 bulls that are in common with method A was slightly decreased with a less stringent data validation.

Conclusions

Data validation is an important part of analysis of producer-recorded health data. Less stringent data validation led to a lower mastitis frequency, indicating that possibly there is some level of underreporting in the Canadian health recording system. Although genetic evaluations stay similar across the investigated data validation methods, future work is necessary to increase data quality in the Canadian health recording system.

Table 4. Pearson correlations between sire breeding values for mastitis resistance from univariate linear sire models, only sires with at least 30 daughters in all data sets were considered (n = 509).

Data validation method	B	C	D	E
A	0.986	0.976	0.957	0.962
B		0.990	0.972	0.973
C			0.984	0.976
D				0.973

Table 5. Number of top 100 bulls in common with data validation method A, only sires with at least 30 daughters in all data sets were considered (n = 509).

Data validation method	Number of top 100 bulls in common with data validation method A
B	91
C	90
D	86
E	85

Acknowledgments

All dairy producers recording health data are gratefully acknowledged. This study was funded by the DairyGen council of Canadian Dairy Network (Guelph, Ontario, Canada) and the Natural Sciences and Engineering Research Council of Canada (Ottawa, Ontario, Canada).

List of References

Carlén, E., E. Strandberg, and A. Roth. 2004. Genetic parameters for clinical mastitis, somatic cell score, and production in the first three lactations of Swedish Holstein cows. J. Dairy Sci. 87:3062-3070.

Dempster, E. R., and I. M. Lerner. 1950. Heritability of threshold characters. Genetics 35:212-235.

Kelton, D. F., K. D. Lissemore, and R. E. Martin. 1998. Recommendations for recording and calculating the incidence of selected clinical diseases of dairy cattle. J. Dairy Sci. 81:2502-2509.

Koeck, A., F. Miglior, D. F. Kelton, and F. S. Schenkel. 2012. Health recording in Canadian Holsteins: Data and genetic parameters. J. Dairy Sci. 95:4099-4108.

Madsen, P., and J. Jensen. 2008. An User's Guide to DMU. A package for analyzing multivariate mixed models. Version 6, release 4.7. Danish Institute of Agricultural Sciences, Tjele, Denmark.

Mrode, R., T. Pritchard, M. Coffey, and E. Wall. 2012. Joint estimation of genetic parameters for test-day somatic cell count and mastitis in the United Kingdom. J. Dairy Sci. 95:4618-4628.

Neuenschwander, T. F.-O, F. Miglior, J. Jamrozik, O. Berke, D. F. Kelton, and L. R. Schaeffer, 2012. Genetic parameters for producer-recorded health data in Canadian Holstein cattle. Animal, 6(4):571-578.

Zwald, N. R., K. A. Weigel, Y. M. Chang, R. D. Welper, and J. S. Clay. 2004. Genetic selection for health traits using producer-recorded data. I. Incidence rates, heritability estimates, and sire breeding values. J. Dairy Sci. 87:4287-4294.