Challenges and opportunities for farmer-recorded data in health and welfare selection

C. Maltecca[1], K.L. Parker Gaddis[1], J. Clay[3] and J.B. Cole[2]

[1]Department of Animal Science, College of Agriculture and Life Sciences North Carolina State University, Raleigh, NC 27695-7621,USA
[2]Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD 20705-2350, USA
[3]Dairy Records Management Systems, Raleigh, NC, USA

Abstract

With an emphasis on increasing profit through increased dairy cow production, a negative relationship with fitness traits such as health has become apparent. Decreased cow health impacts herd profitability because it increases rates of involuntary culling and decreases milk revenues. Improvement of health traits through genetic selection is an appealing tool; however, there is no mandated recording system for health data in the US. Producer-recorded health information provides a wealth of information for improvement of dairy cow health, thus improving the profitability of a farm, yet several challenges remain. The broad definition of 'direct health' does not truly reflect the heterogeneity and complexity of these traits. While there is a virtually endless pool of phenotypes potentially considered for selection, it is paramount to identify a few key parameters for which a consistent and demonstrable improvement can be achieved. We have demonstrated how farmers' recorded events represent a credible source of information with reported incidences matching most of the epidemiological evidence in literature, with calculated incidence rates ranging from 1.37% for respiratory problems to 12.32% for mastitis. Furthermore, we have demonstrated that relationships among common health events constructed from on-farm data provide supporting evidence of plausible interconnection between diseases and overall data quality. The results of our analyses provide evidence for the feasibility of on-farm recorded health base breeding programs. Nevertheless, there is an intrinsic heterogeneity of players, and a complex infrastructure in the collection and flow of information connected to health traits, and among the reasons for the slow implementation of health selection programs, data privacy concerns are at the top of the list in the US.

Keywords: diseases, data validation, US Holsteins.

Introduction

Throughout the past fifty years or more, there has been a focus on increased profit through increasing dairy cow production. With this focus on production, a negative relationship with fitness traits including health and fertility traits, has become apparent (Rauw *et al.*, 1998). An alternative to increasing net profit of producers is to decrease management costs by improving the overall health of the cows. Declining

health of cows can impact the profitability of a herd by affecting several aspects including additional culling, decreased and lost milk, veterinary expenses, and additional labor. Kelton *et al.* (1998) estimated the cost of several common health events ranging from $39 per lactation with an incidence of cystic ovaries up to $340 per case of left displaced abomasum. Over the past fifteen years, however, these economic costs may have drastically changed.

Improvement of health traits by genetic selection is an appealing tool. Difficulty is encountered, however, because there is no mandated or consistent recording system of health traits throughout the United States (Maltecca, 2013). The potential for genetic improvement in health-related traits has been demonstrated in cattle breeds (Abdel-Azim *et al*., 2005; Appuhamy, 2009). Genetic improvement of clinical mastitis incidence has also been demonstrated in Norwegian cattle (Heringstad *et al*., 2003) (more recent). The lack of health-related phenotypes in the US creates an obstacle in achieving genetic improvement of health traits. Several previous studies have confirmed the possibility of using on-farm recorded health information for genetic improvement. (Zwald *et al*., 2004a,b; Miglior, 2009, 2013). In prior research we investigated whether US producer-recorded data reflected the true incidence of health events from epidemiological studies. Further investigation of relationships among occurrences between common health events were compared and corroborated the use of on-farm data as a viable strategy (Parker Gaddis *et al*., 2012).

The use of survey data still poses challenges in terms of data quality and appropriate use. A deeper understanding of causes and distribution for these data is needed. While there is a virtually endless pool of phenotypes that could be potentially considered for selection, there needs to be an effort in identifying a few key parameters for which a consistent and demonstrable improvement can be achieved (Maltecca, 2013). Within this framework an alternative perspective could be used when analyzing health data that aims to extract the underlying health function of a cow. A principal component analysis (**PCA**) may be able to distinguish between groups of health events in order to further elucidate the complex nature of these traits. It could be hypothesized that some cows are more susceptible to a common type of disease, such as reproductive or metabolic due to an underlying disruption in ordinary function. Because of the binary nature of the data, a principle component analysis cannot be directly applied to health event incidence data; however it could be performed on pseudo-phenotypes, such as sire de-regressed breeding values. Alternatively, a multiple correspondence analysis (**MCA**) can be performed directly with binary data (Greenacre and Blasius, 2006). Furthermore, while selecting remains an overarching goal, the nature of disease traits implies a large role in the managing elements of dairy operation, and benchmarking management practices and herd characteristics related to disease incidence can be used to both perform data quality control and risk assessment. In this paper, as part of a larger effort, we provide a preliminary characterization of both individual disease and herd characteristics related to disease incidence.

Material and methods

Data

Voluntary producer-recorded health event data were available from Dairy Records Management Systems (Raleigh, NC) from US farms from 1996 through 2012. The health events included in the analyses were mastitis (MAST), metritis (METR), cystic ovaries (CYST), digestive disorders (DIGE), displaced abomasum (DSAB), ketosis (KETO), lameness (LAME), reproductive problems (REPR), and retained placenta (RETP) from cows of parities one through five. Previous editing was applied to the data for common health events as described in Parker Gaddis *et al.* (2012). Minimum and maximum constraints were imposed on the data by herd-year in order to avoid using records from herd-years that were deemed as either over- or under-reporting. Lactations lasting up through 400 days postpartum were included in the analyses, considering that cows with extended lactations are likely to be those that have not become pregnant.

Table 1. Summary statistics for each health event of interest.

Health event	Number of records	Number of cows	Number of herd-years
Cystic ovaries	222 937	131 194	3 369
Digestive disorders	156 520	97 430	1 780
Displaced abomasum	213 897	125 594	2 370
Ketosis	132 066	82 406	1 358
Lameness	233 392	144 382	3 191
Mastitis	274 890	164 630	3 859
Metritis	236 786	139 818	3 029
Reproductive disorders	253 272	151 315	3 360
Retained placenta	231 317	138 457	2 930

Grouped analyses

Individuals

Several analyses were performed to investigate disease data clustering at an individual level. A MCA was performed using the FactoMineR package (Husson *et al*., 2012) of R (R Core Team, 2012). Many records in the dataset did not have complete observations for all the included health events. The missMDA package (Husson and Josse, 2012) of R was used to impute missing health event observations within the dataset before performing the MCA (Husson and Josse, 2012). A PCA was also performed. Because PCA requires quantitative variables, phenotypes used for this analysis were sire de-regressed estimated breeding values. Estimated breeding values were obtained from a multiple-trait threshold sire analysis using the pedigree-based relationship matrix A. De-regression was performed based on the methodology described by Garrick (2009). The PCA was completed using the FactoMineR package (Husson *et al*., 2012) of R (R Core Team, 2012). Grouped analyses could be considered either from the susceptibility of individuals to certain diseases or with an interpretation stemming from the hypothesis that certain diseases tend to occur together. To determine the optimal number of clusters when considering individual observations, several preliminary analyses were performed. A scree plot was produced to indicate an optimal number of clusters at the inflection point. A hierarchical cluster analysis based on k-means was then performed on the de-regressed sire breeding values as pseudo-phenotypes. The analysis was performed using the fpc library (Hennig, 2013) of R (R Core Team, 2012). Clustering was also performed based on Ward's minimum variance criterion applied to Euclidean distances (R Core Team, 2012).

Herds

Herd summary information was available for four time points throughout each year from 2000 through 2011 for March, June, September, and December. The production, income, and feed cost summary included data such as total number of cows in milk, milk, fat, and protein amounts for the herd, as well as amounts of silage, forage, and concentrates used. The reproductive summary of the current breeding herd included variables such as total number of cows in the breeding herd and voluntary waiting period. A reproductive summary of the total herd included data on percentage of successful services total number of pregnant cows. A stage of lactation profile described data such as number of milking cows by parity group (1st, 2nd, 3+, and all lactations) as well as average daily milk production by parity group. The genetic summary provided data such as the genetic profile and number of service sires used. The production by lactation profile contained descriptive statistics of milk, fat, and protein production split by parity group. A current somatic cell count summary included variables for the percent of cows with a specified SCC level by parity group. The dry cow profile contained the number of days dry for cows in each parity group as well as the number of cows dry for less than 40 days, between 40 and 70 days, and greater than 70 days. Lastly, a yearly summary of cows that entered and left the herd included with data split by parity group.

For this analysis, health information was edited using previously developed criteria to be applied to on-farm recorded data in order to ensure a high quality of the data. The edited health data were then merged with the herd summary data. This resulted in 954,519 records from 266,174 cows across 1,021 herds representing 15,169 sires. A preliminary analysis of the herd variables was conducted in R (R Core Team, 2012) using the caret package (Kuhn, 2013). A function to find any linear dependencies was used to ensure that none of the information provided by the herd summary was completely redundant. After confirming no linear dependencies among the variables, correlated variables were analyzed. The mean overall correlation among the variables was 0.09 with a standard deviation of 0.21. Given that three standard deviation units added to the mean correlation was equal to 0.72, a cut-off for highly correlated variables was designated as 0.75. Variables were removed to minimize the number of highly correlated variables within a dataset. Following this edit, 89 variables pertaining to herd characteristics remained.

Results and discussion

Individuals

In order to investigate how individuals cluster based on their disease liability, hierarchical clustering was performed based on k-means with a k value of 4 based on a scree plot assessment. The results of this analysis are shown in table 2.

In general, the groups tended to be negative for MAST, negative for all events, negative values for metabolic and reproductive events, and positive values for all events. A dendrogram showing the hierarchical clusters based on the pseudo-phenotypes is shown in Figure 1 along with a scatter plot of the individuals using the first two principal components.

When analyzing from the perspective of the health events, LAME and MAST separate very clearly in both the MCA and PCA results. The PCA results indicate that KETO and METR tend to cluster together as well as DSAB and RETP This separation can be seen in Figure 2 showing the variable representation of the PCA. The multiple-trait analysis also estimated a moderate correlation between KETO and METR. The MCA results have a cluster of positive incidences of several health traits including METR, KETO, RETP, DSAB, and DIGE as shown in figure 2. The clear separation of

Table 2. Results of hierarchical clustering based on k-means applied to de-regressed breeding values for sires with estimates for all health events. CYST = cystic ovaries; DSAB = displaced abomasum; KETO = ketosis; LAME = lameness; MAST = mastitis; METR = metritis; RETP = retained placenta.

Cluster	CYST	DSAB	KETO	LAME	MAST	METR	RETP
1	-0.31	1.27	1.06	-0.01	-0.19	0.58	0.42
2	0.22	-0.62	-0.40	0.08	-0.09	-0.14	-0.11
3	0.60	-2.16	-1.33	0.16	0.09	-0.67	-0.53
4	0.02	0.30	0.18	-0.05	0.05	0.06	0.01

CYST in the PCA is not seen in the MCA results. However, if imputed data is not included when performing the MCA, the separation of CYST is observed. This may indicate that the separation of CYST in the PCA, as well as the MCA without imputing missing data, is an artifact resulting from having incomplete records. In general, the MCA and PCA results indicate that several of the health events do tend to cluster together. This indicates that there is the possibility of creating broad health event definitions while not losing a large amount of information. For example, based on biological knowledge as well as the MCA and PCA results, groups of events could be formed for mastitis or other udder-related disorders, lameness and foot or leg problems, reproductive disorders, and metabolic disorders. This reduces the details that are needed from producers while still allowing informative health data to be collected.

Herds

A principle component analysis was performed on the 89 herd variables to determine if certain characteristics tended to occur together. Eleven components explained about 50% of the total variation explained by the herd variables while twenty-eight components explained 75% of the total variation. A description of the dimensions was also inspected. Based on the results, somatic cell counts in first and second lactation cows and the number and average age of cows across all lactations were the most characteristic of the first dimension. The second dimension most highly reflected production traits such as rolling average of milk pounds, summit milk of first lactation cows, average daily milk production from 1 to 40 DIM for all cows, and fat yield. The third dimension reflected the number of cows dry over seventy days, the number of cows entering the herd, and the number of cows dry less than forty days.

Herd variables were clustered in regard to the crude incidence of common health events. Each health event was analyzed individually. The optimum number of clusters for the data was estimated and observations were split into the optimum number of clusters "around medoids". For each event, the optimal number of clusters was two. Following clustering, the average of select herd characteristics are given for each health event in table 3. The number of second lactation cows entering the herd was greater in the cluster with lower incidences of mastitis, ketosis, and retained placenta. Herd characteristics that involved number or percentage of cows leaving the herd were among the characteristics that were most different between the clusters of herds for mastitis, metritis, ketosis, and retained placenta. The herds

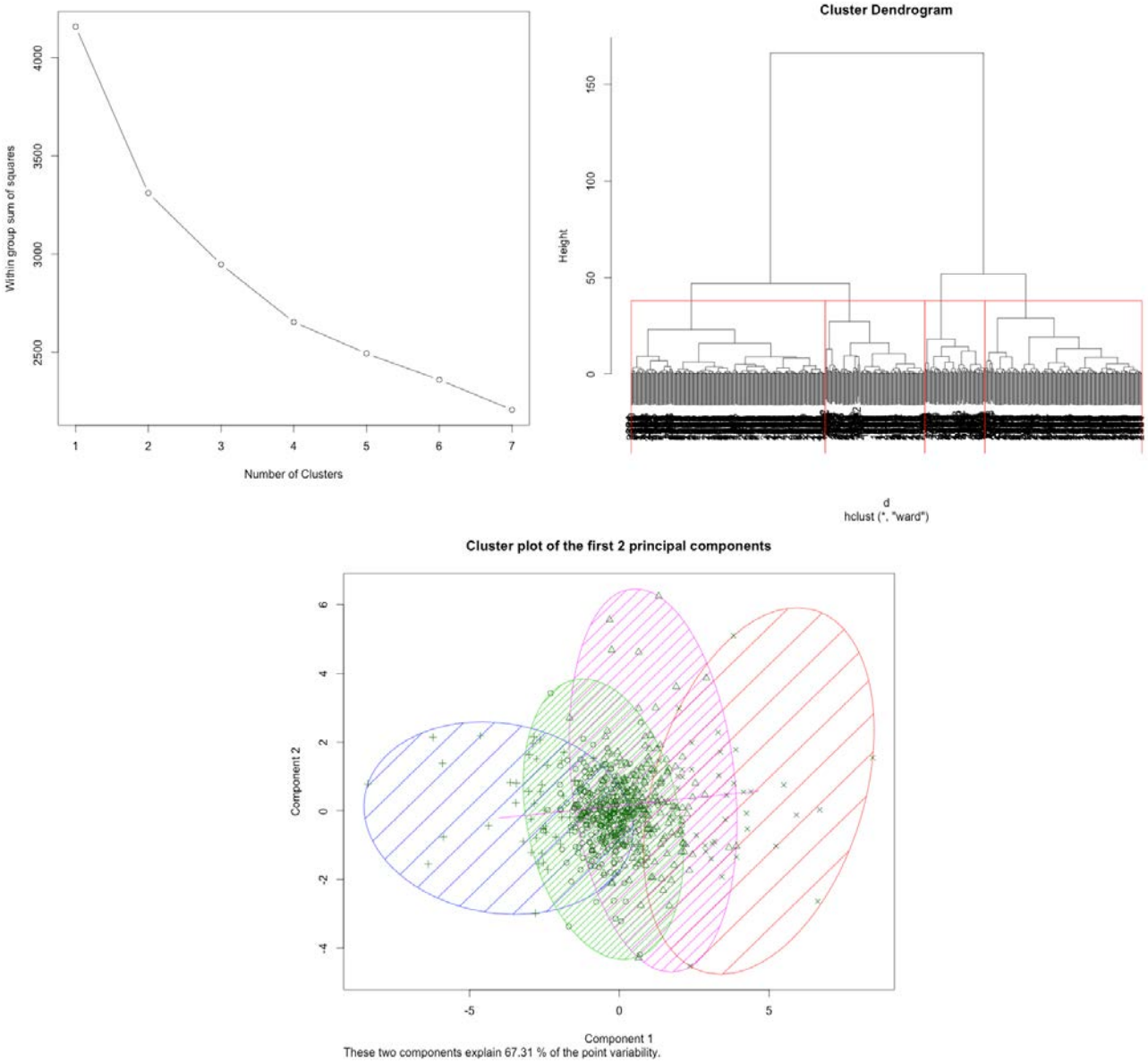

Figure 1. Panel A: Scree-like plot indicating within-group sum of squares with each number of clusters using scaled de-regressed sire estimated breeding values as pseudo-phenotypes. Panel B. Cluster dendrogram showing hierarchical clustering results using de-regressed sire breeding values as pseudo-phenotypes for sires with estimates for all health events. Red rectangles indicate the optimal four clusters. Panel C Cluster plot of individuals against the first two principal components determined based on pseudo-phenotypes of de-regressed sire estimated breeding values for all health events. Each cluster is shown by an ellipse with each individual depicted by either a circle, square, cross, or x.

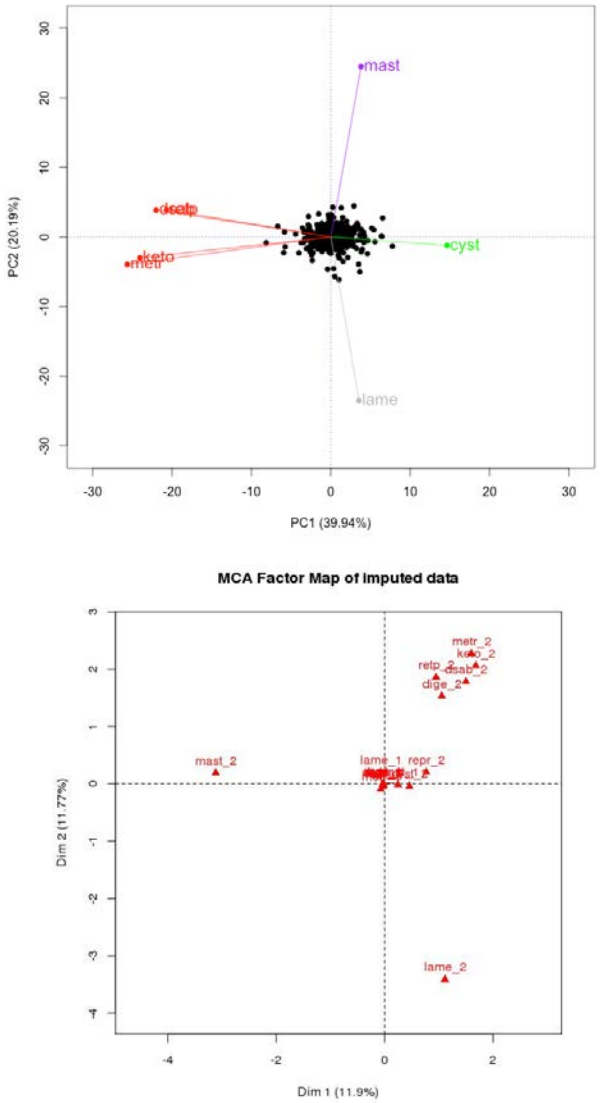

Figure 2. Panel A: Principle component analysis variable factor map using de-regressed sire estimated breeding values for sires with estimates for all events. The first two components (Dim 1 and Dim 2) are displayed, explaining 39.94% and 20.19% of the variance, respectively. CYST = cystic ovaries; DSAB = displaced abomasum; KETO = ketosis; LAME = lameness; MAST = mastitis; METR = metritis; RETP = retained placenta. Panel B: Multiple correspondence analysis factor map from imputed data portraying the first two dimensions (Dim 1 and Dim 2) which explain 11.9% and 11.77% of the total variance, respectively. No incident = 1, Incident = 2. CYST = cystic ovaries; DIGE = digestive disorders; DSAB = displaced abomasum; KETO = ketosis; LAME = lameness; MAST = mastitis; METR = metritis; REPR = reproductive disorders; RETP = retained placenta.

clustered with lower incidences of the analyzed diseases had fewer cows leaving the herd, whereas herds clustered with high incidences had a greater number of cows leaving the herd. Herds that were clustered in the low incidence groups for metritis, mastitis, and ketosis all reported having higher numbers of total cows and milking cows on pasture. Total number of services was a herd characteristic that

Table 3. Average of select herd characteristics based on clustering results.

Health Event	Incidence (Group)	Total cows (RA)	Milk lbs. (RA)	Fat lbs. (RA)	Protein lbs. (RA)	Avg. days to 1st service	Actual calving interval	Avg. % successful services	Total number calving	Avg. daily milk production	Body weight
RETP	0.005 (Low)	333	20884	792	653	84	13.9	36	561	66.6	1270
	0.10 (High)	444	22113	834	687	82	13.8	34	892	70.1	1306
MAST	0.008 (Low)	263	20872	794	652	85	14.0	36	349	66.5	1273
	0.16 (High)	554	21269	805	663	83	13.9	36	1110	68.3	1286
METR	0.01 (Low)	322	21026	803	660	86	14.0	35.3	445	66.8	1274
	0.14 (High)	578	21582	813	667	83	14.0	34.5	707	69.0	1310
KETO	0.006 (Low)	370	21744	833	681	79	13.8	32.7	427	69.3	1270
	0.10 (High)	441	22569	853	701	78	13.8	31.3	682	72.5	1311

RA = rolling yearly herd average; RETP = retained placenta; MAST = mastitis; METR = metritis; KETO = ketosis.

Table 4. Herd variables with greatest relative importance by health event.

Health event	Variable 1	Variable 2	Variable 3
Ketosis	Milk yield (all cows)	Avg. days to 1st service (2nd lactation cows)	Services per pregnancy (pregnant 1st lactation cows)
Mastitis	Average total pregnant cows	Voluntary waiting period	Total cows
Metritis	Voluntary waiting period	Total cows	Feed cost per cwt milk
Retained placenta	Average total pregnant cows	Average percentage heats observed	Pounds concentrate consumed

was among those most different between the two clusters for several diseases. Herds with low incidences of retained placenta, ketosis, or mastitis had less total number of services than herds with high incidences of those diseases.

Conclusions

Opportunities exist to improve disease prediction and overall herd disease management by making use of patterns observed at both individual and herd level. Grouped information can be used in data editing and herd benchmarking, as well as a way to increase selection efficacy. Further evaluations of more comprehensive predictive models are nonetheless required.

List of References

Abdel-Azim, G.A., A.E. Freeman, M.E. Kehrli, S.C. Kelm, J.L. Burton, A.L. Kuck, and S. Schnell. 2005. Genetic basis and risk factors for infectious and noninfectious diseases in US Holsteins. I. Estimation of genetic parameters for single diseases and general health. J. Dairy Sci. 88:1199-1207.

Appuhamy, J.A.D.R.N., B.G. Cassel, and J.B. Cole. 2009. Phenotypic and genetic relationships of common health disorders with milk and fat yield persistencies from producer-recorded health data and test-day yields. J. Dairy Sci. 92:1785-1795.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Greenacre, M., and J. Blasius. 2006. Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton.

Hennig, C. 2013. fpc: Flexible procedures for clustering. R package version 2.1-5.

Husson, F., and J. Josse. 2012. missMDA: Handling missing values with / in multivariate data analysis (principal component methods). R package version 1.5.

Husson, F., J. Josse, S. Le, and J. Mazet. 2012. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.20.

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.

Kuhn, M. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, and T. Cooper. 2013. caret: Classification and Regression Training. R package version 5.15-61.

Heringstad, B., Y.M. Chang, D. Gianola, and G. Klemetsdal. 2003. Genetic analysis of longitudinal trajectory of clinical mastitis in first-lactation Norwegian cattle. J. Dairy Sci. 86:2676-2683.

Maltecca, C. 2013. Fitter happier: the never-ending quest for a better cow. J. Anim. Breed. Genet. 130:87-88.

Parker Gaddis, K.L., J.B. Cole, J.S. Clay, and C. Maltecca. 2012. Incidence validation and relationship analysis of producer-recorded health event data from on-farm computer systems in the United States. J. Dairy Sci. 95:5422-5435.

R Core Team. 2012. R: A language and environment for statistical computing. *www.r-project.org*

Rauw, W.M., E. Kanis, E.N. Noordhuizen-Stassen, and F.J. Grommers. 1998. Undesirable side effects of selection for high production efficiency in farm animals: a review. Livest. Prod. Sci. 56:15-33.

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

Zwald, N.R., K.A. Weigel, Y.M. Chang, R.D. Welper, and J.S. Clay. 2004b. Genetic selection for health traits using producer-recorded data. II. Genetic correlations, disease probabilities, and relationships with existing traits. J. Dairy Sci. 87:4295-4302.