Processing of data discrepancies for U.S. dairy cattle and effect on genetic evaluations

G.R. Wiggans & L.L.M. Thornton

Animal Improvement Programs Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA

Summary

Genetic evaluations depend on accurate data for

1. Accuracy of the recorded trait.
2. Information on how much emphasis the trait should receive and how it should be adjusted.
3. Which other animals the trait should influence.

Errors in any of those areas reduce accuracy of all evaluations. When problems are detected, data are rejected or modified to remove the inconsistency. Annotated records that indicate rejection or change are returned to processing centers for review by their personnel to assist in data correction and to explain actions taken. With increasing computing power, reducing the number of data errors and discrepancies should be possible, and error reports and correction should be simpler.

***Key words**: Milk recording, Data conflicts, Dairy herd improvement.*

Introduction

Data integrity and verification are important parts of data utilization (Decker and Martinenghi, 2006). Particularly with multiple contributors, care must be taken to ensure that data are accurate and remain intact when updates are made (Karpovsky *et al.*, 2003; Smith *et al.*, 2005).

The Animal Improvement Programs Laboratory (AIPL) calculates genetic evaluations of U.S. dairy cattle for yield traits, longevity (productive life), mastitis resistance (somatic cell score), fertility and calving traits of dairy cattle as well as conformation (type) traits for some breeds. Those evaluations depend on accurate data from several sources. Five dairy records processing centers provide milk recording data. Breed registry societies provide pedigree and type data. Data on calving traits and bull status for artificial-insemination service are provided by the National Association of Animal Breeders (NAAB). Extensive data checking is done to enforce consistency. For example, errors and changes in animal identification (ID) can cause data from 2 animals to be assigned to the same animal or data from the same animal to be treated as 2 animals. Formatting or entry errors can result in reporting of impossible values; however, unusual values that are correct must be allowed.

Genetic evaluations depend on accurate data for:

1. Accuracy of the recorded trait (e.g., milk weight).
2. Information on how much emphasis the trait should receive and how it should be adjusted (e.g., milking frequency, number of milkings weighed).
3. Which other animals the trait should influence (e.g., parents, progeny and contemporaries) (Wiggans *et al*., 2007).

Errors in any of those areas reduce accuracy of all evaluations. Banos *et al.* (2001) documented the effect of misidentification on accuracy of genetic evaluations. As genomic information becomes more prevalent and is used in genetic evaluations, data integrity becomes more important, and additional checking must be added to verify accuracy of both traditional phenotypic data and genomic information (Kemmeren *et al.,* 2002).

Data flow

For national genetic evaluations of the U.S. dairy population, pedigree and yield records are processed by the main editing program, which checks pedigree and yield fields before records are added to the AIPL master database. For new animals, their ID is checked to see if it is valid. A valid ID includes a valid breed, country code and ID number. Canadian animal ID is validated against a list of ID supplied by the Canadian Dairy Network (Guelph, ON). For 9-digit American ID, the last digit is a check digit that can detect some invalid ID. An animal's birth date then is checked against its parents' birth dates to see if the sire or dam was either too young or too old to be a parent. If a dam has lactation records, her progeny's birth date is matched with her calving dates. Differences of <1 month are allowed. If an animal is from embryo transfer, comparison with dam calving date is omitted. Parents without information in the database are added with an estimated birth date that is assigned to be 3 years before the reported animal's birth date. Estimated dates are revised if information from older siblings is received.

Detection of additional ID (aliases) for an animal is complex. Animals that have the same birth data and are full siblings but not twins are investigated. The within-herd ID (control number) is valuable in determining whether the ID is for a new animal or an alias. Bulls registered in >1 country have been a common cause of an animal's being treated as >1 animal. That problem has subsided as countries have accommodated foreign ID and have stopped re-registering bulls. Identification numbers that differ by only 1 digit (either different or missing) are investigated as possibly invalid ID for the same animal. For cows with lactation data, yield data must not conflict for data from 2 ID to be combined as data for the same cow.

Yield data are edited based on 2 ranges. Values outside the widest range are rejected as invalid. Values outside a more narrow range determined for each cow's lactation are stored; however, when they are used, they are changed to a floor or ceiling derived from other test-day data. Careful consideration is given to determining which herd a cow was in when she produced a test-day record. A herd test-date record identifies valid herd test dates. Otherwise, cows with a test day in a previous herd would have those data assigned to the current herd. Checking also is done to assign events to the correct calving date, which most often is an issue with breeding information. A missing calving date causes such events to be associated with the previous calving.

Error reports

When an error or conflict is detected, a record that has been annotated to indicate an editing action of reject, notify, or change is returned to the processing center to assist in data correction and to explain what action was taken. Typically, those records are stored to assist in answering queries and, in some cases, forwarded to the supervisor or producer for action. Records that are rejected also are available by query on the AIPL website (*http://aipl.arsusda.gov*) to assist in answering questions and correcting problems. Table 1 shows common errors and their frequencies for 1 day's edits of pedigree records from 1 breed registry society and lactation records from 1 dairy records processing center.

Table 1. Frequency of common errors in pedigree and lactation records submitted on 1 day from 1 breed registry society and 1 dairy records processing center for addition to the U.S. database of records for calculation of genetic evaluations for dairy cattle.

Record type	Error code	Definition	Disposition	Frequency (no.)
Pedigree (n = 12 000)	1Nd	Merging input to animal in master	Notify	207
	1Oh	Update input to twin	Change	138
	3Ib	Dam identification (ID) differs from master, source not verified	Notify	107
	1Od	Sibling updated to twin	Notify	106
	2Be	Sire ID not preferred	Change	82
	3Be	Dam ID not preferred	Change	75
	5Fc	Birth date and dam calving date not the same	Notify	69
	2Jc	Sire ID differs from service sire ID	Notify	64
	2Ib	Sire ID differs from master, source not verified	Notify	60
	4Jc	Master same as cross-reference	Change	52
Lactation (n = 93 000)	2De	Grade sire misidentified	Reject/change	2 738
	1Be	ID not preferred ID	Change	2 611
	6Td	Parity and age mismatched	Change	2 334
	0Jd	Multiple birth code ignored	Change	2 235
	7Ic	Abnormal recorded milk yield	Change	1 967
	2Gd	Sire ID differs from master	Change	1 902
	7Ob	Quality control code incorrect	Notify	1 899
	5Bd	Birth date differs from master	Reject	1 801
	3Gd	Dam ID differs from master	Change	1 707
	7Mb	Milkings weighed not the same as for herd	Change	1 472

Editing principles

When errors are encountered, data are either rejected or modified. Data rejection causes loss of possibly valuable information and often results in no genetic evaluation for animals of interest. Therefore, the system is designed to retain data whenever possible. However, data elimination is preferred to retention of conflicting data. For example, if an animal's birth date conflicts with its dam's calving date and both animals already have data in the system, the dam ID is removed to resolve the conflict and to allow records for both animals to remain in the database.

Importance of types of data

Milking times

To save labor costs, most U.S. herds are enrolled in an a.m. - p.m. test plan in which not all milkings are supervised and daily yield is estimated from the recorded milking. That estimation is based on the interval since the previous milking. Although the start time for each milking is critical, milkings vary in length (e.g., supervised milkings generally are longer). A more accurate estimate of the interval between milkings can be derived from the midpoints of consecutive milkings instead of starting times, which is why an end time also is required for each milking.

Alternation of supervised milking

The purpose of milk recording is to estimate yield accurately for a dairy. Because each dairy is unique, a.m.-p.m. estimation formulas derived for the national population are not an exact fit for individual dairies. If supervised milkings alternate over time between morning and evening, systematic errors should average out, and lactation records would be unbiased. Such alternation of supervised milkings sometimes is difficult to achieve with large herds.

Herdmate ID

Genetic evaluations rely heavily on pedigree data, particularly so that bulls of superior genetic merit can be identified. Therefore, data from cows with unknown sires are not included in evaluations. Only evaluated cows can serve as herdmates of other cows; thus, even large herds may have small contemporary groups if most cows are not sire identified.

Breed reporting for crossbreds

All breeds are included in the U.S. across-breed genetic evaluation (VanRaden *et al.*, 2007), and the breed percentages for an animal are derived from its pedigree. The breed determines the breed base on which a cow's evaluation is reported unless the breed is coded as XX (crossbred). Sire breed determines the breed base for evaluations of crossbred cows. Generally, an animal's breed should reflect the breed with the highest percentage from within the animal's pedigree. For crossbred herds, genetic evaluations are likely to be reported on different breed bases. For animals with equal breed percentages (e.g., 50% Holstein, 50% Brown Swiss), using the predominant breed for the herd is beneficial.

Data collection rating

The data collection rating (DCR) measures how much data from a particular test plan are expected to vary from a standard. The less information that is collected, the lower the DCR. However, DCR does not measure bias directly. For example, if the same milking is sampled every month in a herd enrolled in a.m.-p.m. testing with component sampling, the estimates of component yields will be biased by the degree that national estimation formulas do not fit the herd. However, the amount of

information collected is not different; thus, the variance of the error is not increased, and the DCR is the same. The DCR for unsupervised milkings is arbitrarily set to 75% of that for a supervised milking. A similarly discounted DCR could be used for herds enrolled in a.m.-p.m. testing when the sampled milking is not alternated if that information were reported to AIPL.

Automatic milk recording

Automatic milk recording (AMR) equipment provides an opportunity for increased recording accuracy. However, AMR systems must monitor their own accuracy and detect when a unit needs maintenance. Because they depend on accurate cow ID, they can be subject to misread ID and scrambling of cow order while filling stalls. Typically 5- to 10-day averages are reported, which limits the effect of assigning an individual milking to the wrong cow. The AMR system should detect atypical cow yields and exclude them. The editing system for the AIPL database cannot detect common AMR problems; thus, accurate meter calibration is important.

Conclusions

The AIPL system used for checking data used in national U.S. genetic evaluations of dairy cattle is highly complex. Records from various sources are combined, and conflicting data are harmonized based on which data are expected to be most accurate. Conflicting data are deleted when necessary. Data for pedigree, yield and animal status all affect genetic evaluations, which can only be as accurate as the contributing data. Invalid records can diminish the accuracy of evaluations for other animals. Accurate data are most likely to be generated if those collecting it understand how the information is used and how accurate information benefits the entire data scheme.

Acknowledgements

The authors acknowledge the contributions of John Clay at Dairy Records Management Systems (Raleigh, NC), Dan Webb at the University of Florida (Gainesville, FL) and Lillian Bacheller at AIPL (Beltsville, MD).

List of references

Banos, G., G.R. Wiggans & R.L. Powell, 2001. Impact of paternity errors in cow identification on genetic evalutions and international comparisons. J. Dairy Sci. 84: 2523–2529.

Decker, H. & D. Martinenghi, 2006. Avenues to flexible data integrity checking. 17th Inter. Conf. on Database and Expert Systems Applications (DEXA'06), pp. 425–429.

Karpovsky, M.G., L.B. Levitin & A. Trachtenberg, 2003. Data verification and reconciliation with generalized error-control codes. IEEE Transactions on Information Theory 49: 1788–1793.

Kemmeren, P., N.L. van Berkum, J. Vilo, T. Bijma, R. Donders, A. Brazma & F.C.P. Holstege, 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Molecular Cell 9: 1133–1143.

Smith, A., D. Greenbaum, S.M. Douglas, M. Long & M. Gerstein, 2005. Comment: Network security and data integrity in academia: an assessment and a proposal for large-scale archiving. Genome Bio. 6: 119.

VanRaden, P.M., M.E. Tooker, J.B. Cole, G.R. Wiggans & J.H. Megonigal, Jr., 2007. Genetic evaluations for mixed-breed populations. J. Dairy Sci. 90: 2434–2441.

Wiggans, G.R., M.A. Faust and F. Miglior, 2007. Harnessing automatic data collection to enhance genetic improvement programs (abstract). J. Dairy Sci. 90 (Suppl. 1): 377.