VALIDATION OF DATA AND REVIEW OF RESULTS FROM GENETIC EVALUATION SYSTEMS FOR US BEEF AND DAIRY CATTLE

J.K. Bertrand[1] and G.R. Wiggans[2]

[1]Animal and Dairy Science Department, University of Georgia, Athens, 30602-2771, USA
[2]Animal Improvement Programs Laboratory, Agricultural Research Service,
US Department of Agriculture, Beltsville, MD 20705-2350, USA

SUMMARY
Estimates of genetic merit are used as both selection and marketing tools and, therefore, should predict the merit of future progeny as accurately as possible. Success depends on having an appropriate model and adequate data. The goal of data editing is to exclude questionable information from genetic evaluations so that the evaluations are as accurate as possible while still remaining representative of the population. Proper editing and contemporary group formation can protect against some errors in data reporting. However, any exclusion of data must be able to be justified. Although many data errors affect individuals rather than the entire population, minimization of errors is critical to building confidence in evaluations. A few individuals with high estimates of genetic merit that are not warranted can erode trust in a genetic evaluation system.
Keywords: data validation, genetic evaluation, beef cattle, dairy cattle

INTRODUCTION
Genetic evaluation systems are widely accepted as a means of estimating genetic merit of breeding animals. They usually are defined in terms of the models on which they are based. The genetic parameters used indicate likely response to selection and accuracy of estimated breeding values.

Another key component of a sound genetic evaluation program is pedigree and performance data that are free of major errors that could bias predicted genetic values or reduce their accuracy. Studies by Van Vleck (1970) with dairy cattle data, Christensen *et al.* (1982) with dairy and dual-purpose cattle data, and Long *et al.* (1990) with swine data reported that reranking of evaluated animals occurred as a result of pedigree errors and animal misidentification. Mallinckrodt *et al.* (1992) used simulation data to study loss of reliability in prediction of breeding values due to several different types of data problems. Significant loss of reliability was observed as a result of selective reporting or misreporting of performance records, misrepresentation of contemporary groups, and misidentification of sires.

General protocols follow for editing of performance and pedigree data and review of results from current genetic evaluation programs for US beef and dairy cattle.

BEEF CATTLE
Merchandising of yearling and 2-year-old seedstock, particularly bulls, is prevalent in the beef industry in many parts of the world. For these young animals, a combination of pedigree information and the animal's own performance record often are the only data available to estimate a breeding value to be used for selection and merchandising purposes. Because of the importance of individual performance records in prediction of breeding values for young animals, valid data editing and contemporary group formation are critical so that genetic values can be used with confidence by producers.

In many countries, data used for genetic evaluation of beef cattle are collected directly by producers with little or no oversight by any government or breed organization. Beef cattle breed associations provide varying amounts of performance and pedigree information for genetic evaluation, a reflection of the lack of rigid beef industry standards for data collection. Estimates of animal misidentification by the US beef industry based on breed associations that randomly blood test are between 5 and 10%. Between 25 and 55% of data submitted by US producers is eliminated prior to genetic evaluation because of data errors.

Edits and data validation checks for the US beef evaluation system include animal identification and pedigree validation, performance record validation, and contemporary group validation.

Animal identification and pedigree validation. Identification numbers are checked to determine if they fall within an acceptable range. Pedigrees are checked to ensure that individuals are not in their own ancestry and to

ensure that individuals are not in pedigrees as both a sire and a dam. Depending on the type of files provided by the breed association, data are examined to locate inconsistent pedigree information in the performance record file or to locate pedigree information in a separate pedigree file that does not match information contained in the performance file. When inconsistencies are encountered, records must be eliminated or corrected depending on the seriousness of errors.

Animal identification and pedigree integrity have taken on new significance as breed associations in Canada and the United States merge their data bases to compute joint genetic analyses. Unique animal tattoo numbers are not maintained across countries, and the registration number in the country of origin does not always remain with the animal as it is used in the other country. Therefore, combinations of birth dates, registration numbers, tattoo numbers, names, and parental information are used to identify animals across countries. Data files for sires that have been identified as being the same animal in Canada and the United States and sires that have a high probability of being the same animal in both countries are sent to breed associations for additional verification.

Birth dates of animals and parents are checked to ensure that animal age at recording and parent ages at birth of progeny are reasonable.

Many breed associations allow cattle that are less than 100% purebred to be registered and provide a prefix to the registration number to indicate breed percentage or breed-type status of the animal. Because breed percentage or type often is used to form contemporary groups or to account for breed group or heterosis differences in the analysis, the prefix to the registration number is checked against the percentage designation of the animal on the performance record. Some breed associations allow performance records from animals with unidentified dams, whereas others require that both parents be identified and registered.

Performance record validation. To avoid incorrectly entered or highly suspect data, records are excluded if they have animal ages or measured traits that are outside breed-specific ranges. Ranges for animal age are chosen so that retained records can be adequately adjusted for age in the analysis. Trait ranges are upper and lower limits of what is possible.

After contemporary groups are formed, mean of each contemporary group is computed, and a ratio is calculated for each record relative to contemporary group mean. Records within a group are eliminated if they are outside the ratio range of 60 to 140%, which usually is equivalent to eliminating records that differ from the mean by more than 3.5 to 4 standard deviations. This edit is an attempt to eliminate records from animals that may have been sick, preferentially treated, or placed in the wrong contemporary group.

Progeny records are excluded if a dam is younger than 18.3 months (550 days) or older than 20 years (7300 days) at calving. Postweaning records are eliminated if weaning information is not available for the animal.

Contemporary group formation and validation. Improper contemporary group formation or incorrect assignment of animals to contemporary groups can result in biased breeding values. For example, for one US bull, predicted breeding value for milking ability decreased more than 2.5 standard deviations in consecutive evaluations. The bull had 130 daughters in production for the first evaluation and 57 additional daughters for the next. Examination of data revealed extremely negative contemporary group deviations for grandprogeny of the bull from a breeder who had reported information for 44 of the new daughters that came into production. This breeder had two farms in two different states, and daughters from the bull of interest were located on the farm with the most limiting environment. The breed association provides only a breeder code and no herd or farm code. Calves at both farms were weighed on approximately the same date and, therefore, appeared to be in the same contemporary group. The breeder should have sent information from each farm on separate performance sheets with different invoice numbers because invoice processing number is included in definition of contemporary groups for the breed. However, the breeder pooled data from both farms on one performance sheet. After the mistake was corrected and data reanalyzed, the bull's new predicted breeding value for milk was similar to the one from the previous evaluation.

Information that is available to form contemporary groups varies by breed. The minimum information needed to form a birth contemporary group is herd, calf gender, and season or 90-day group. In addition to these three basic criteria, lot processing date or performance invoice processing number, breeder-defined pasture or group code, dam breed, and breed percentage for the calf have been used. For weaning (205 or 200 days) contemporary group, herd, calf gender, and weaning weigh date are the basic criteria for group formation. Other information that could be

used include weaning management code (creep or noncreep), weaning lot processing date or weaning performance invoice processing number, breeder-defined pasture or group code, dam breed, and breed percentage for the calf. Basic criteria for formation of yearling (365, 452, or 550 days depending on country) contemporary groups are herd, calf gender, weaning contemporary group as previously defined, and yearling weigh date. Additional information that has been used is yearling management code, yearling lot processing date or yearling performance invoice processing number, breeder-defined group, dam breed, and breed percentage for the calf.

Certain checks are conducted to gauge correctness of a contemporary group. If all birth weights within a group are equal, that group is excluded. For traits for which a weigh date or processing date is used to form the contemporary group, ages of the animals are checked to ensure a reasonable range of animal ages within the group. Data for individual animals or entire contemporary groups may be eliminated if inconsistencies in animal ages are encountered.

For many beef cattle herds, records are not submitted from all calves in a contemporary group. Garrick *et al.* (1989) reported that 63 and 68% of birth and weaning weight records, respectively, available for Simmentals were from female calves. Herds with only a portion of records from their male calves can be identified by checking their male-to-female ratio for performance records submitted over time. Elimination of performance records for male calves from these identified herds could reduce bias in predicted breeding values due to preferential reporting. Another possible method for reducing bias that results from preferential submission of records is to include a random effect for interaction of sire and contemporary group in the analysis model. Further research is needed to define optimal methods to account for preferential reporting in evaluation systems.

Direct connectedness of sires across herds and contemporary groups is checked for most breeds. Percentage of sires that are disconnected from the main group of sires based on direct sire connectedness across contemporary groups can be as high as 10%. If contemporary groups that contain progeny from disconnected sires are retained when at least one of the sires of animals in the contemporary group is a sire or a son of a sire in the main connected body of data, the maximum percentage of sires that are considered disconnected for any breed is less than 5%. Progeny performance records from these disconnected sires are eliminated prior to genetic analysis for those breeds where connectedness is checked.

Review of results. Information is collected throughout analysis to provide quality control. Numbers of records and effects are compared with the previous analysis to ascertain that both have increased as expected. In addition, breeding values are checked against those from the previous analysis to determine if all animals with breeding values computed in the previous analysis have breeding values in the current analysis. When discrepancies are encountered, analysis is suspended until the cause of the loss of information is ascertained and corrected if appropriate. Means and standard deviations for breeding values from current and previous analyses are compared, and further investigation is conducted if large changes in these values have occurred. Rank correlations between breeding values computed in current and previous analyses also are checked to ensure that rank correlations are equal to or greater than those generated with the previous analysis.

Animals with the largest changes in breeding value between previous and current analyses are identified, and a list of these animals is sent to breed associations along with a list of animals that were eliminated because of edits. For each breed, one or two sires that had the largest change in breeding value but greatest accuracy for predicted breeding value from the previous analysis are investigated further. Other comparisons, such as progeny contributions of the sire with other sires across herds and contemporary groups or breeding values of mates, parents, and progeny, are used to investigate reasons for changes in breeding values. Occasionally, contemporary grouping errors are found when investigating changes in sire breeding value because large progeny deviations occur in contemporary groups from one herd. In response to breeder concerns, breed associations frequently request that breeding values be examined for specific sires and cows with significant changes in breeding value.

DAIRY CATTLE

Genetic evaluations long have been of primary importance in determining economic and breeding merit of dairy cattle, particularly for semen and embryos. The first US milk recording association was begun in 1906, and genetic evaluations were computed and released to the industry starting in the mid 1930s. Until 1997, genetic evaluations for yield traits included only data collected under the supervision of a third party. In February 1997, milk yields recorded by owners were included in US evaluations but were subject to more stringent editing.

Although the testing supervisor traditionally had been relied on to ensure data quality in the United States, the Dairy Herd Improvement program was reorganized in January 1997 to allow a reduction in rule enforcement. Responsibility for monitoring accuracy of recorded data has been shifted to users, and a herd profile was developed so that users can judge data quality. Herd profiles include details about the test plan, including when tests occurred, if they were supervised, and which milkings were weighed and sampled. These profiles also provide identification of data outliers. A data collection rating has been developed to measure the amount of information included in a lactation record. This rating is based on the expected correlation between lactation records with the information recording characteristics of this record and records calculated from 10 equally spaced tests and samples.

Changes in the US record-keeping program place further responsibility on those who compute genetic evaluations to determine which data are suitable. Many checks have been imposed to ensure consistency and reasonable values. The editing system has 9 categories of errors and 87 individual codes, some of which cover several problems for the same error. Three dispositions of records result from discovery of an error: rejection, modification, and notification.

Pedigree edits. For genetic evaluations to discriminate among animals, performance data must be assigned to the correct animal and relationships among animals must be accurately reported. Frequently, animals are reidentified. Eartags may be lost, or an animal may be reported with an eartag and then receive a registration number from a breed association; the result is that data are reported under both identification numbers. Parent identification may be wrong so that progeny contribute to the wrong animal. The greater the degree of misidentification, the greater the difficulty in detecting the outstanding animals because effective heritability is lowered.

For US yield evaluations, pedigree data are checked (Norman *et al.* 1994) to determine if identification numbers fall within appropriate ranges by breed. Animals must be born at least 14 months after their parents, and an animal's birth date must match calving date of the dam unless the animal results from embryo transfer. Possible reidentification is detected when two animals have the same parents and birth date unless they are twins or result from embryo transfer. To give flexibility to the edit system so that unusual situations can be accommodated, edits can be overridden using a verification code. When conflicts occur between pedigree data provided by a breed association and a dairy records processing center, data from the breed association usually are retained.

Notification of data changes or rejection is returned to US dairy records processing centers or breed associations that supply the data (and in some cases to artificial-insemination organizations) for review and correction. Dairy record processing centers relay these error records to their producers. Notification records are useful in explaining why an animal was not evaluated. Dairy records processing centers often send a record for the first test day for each cow in first lactation so that pedigree data can be checked and corrected before the next evaluation is computed.

The greatest portion of US dairy data is excluded because of missing sire identification. A system has been developed to enable collection of pedigree data nationally from calf records so that information will be available for animals that are sold as heifers. A national identification system with a single series of numbers across breed and gender will be implemented in 1998. The new system should solve reidentification problems except when animals lose their identification tags.

Bull owners receive a list of bull daughters with a bull's evaluation. This list enables detection of cows that could not be daughters of a bull because of the date of semen release. Artificial- insemination organizations make an important contribution to ensuring that data for bull daughters are corrected.

Yield edits. As with pedigree records, US yield records with unusual values are excluded. Limits are placed on mean milk, fat, and protein yields per day and fat and protein percentages. These limits vary by breed and parity (first parity versus later parity). Calving dates are required to be at least 9 months apart.

A problem in evaluating dairy cattle is that cows change herds; therefore, determination of appropriate contemporaries can be difficult. For US genetic evaluations, data only from the first herd are used unless the number of days in milk reported for the first herd are fewer than 90 and also less than half the number of days in milk reported for the second herd. For the test-day model currently being developed for US evaluations, yields will only be compared to those of other cows in the same herd milking on the same day.

Additional edits are imposed for US data from herds sampled by their owners rather than a testing supervisor. These edits include a reasonable correspondence between bulk tank measurements and the sum of the individual

cow milk weights (milk weights must be more than 80% and less than 118% of milk shipped), 40% valid identification for animals in the herd and their sires and birth dates, and a proportion of outlier records in the herd below a threshold based on herd size. Determination of outliers is based on interquartile ranges.

For US dairy cattle, contemporary groups (called management groups) are defined with the goal of including five lactation records. Management groups are separate for first and later parities and for herdbook-recorded cows. If fewer than five records are available for a 2-month period, adjacent groups are combined to include up to 6 months, and the distinction for herdbook registry is eliminated. If at least five records still are not available, records from first and later parities are combined, and target size for the management group is reduced to 3 lactation records. To reach the 3-record target, groups are combined further to include a maximum of 12 months. The definition of management groups is based on herd code. However, herd codes can be changed, and consequently cows that should be in the same management group sometimes are not. Contemporary grouping for US dairy cattle could be improved by identifying strings of cows within a herd that are managed differently. Because data collection for dairy records is frequent and traditionally has been supervised, formation of inappropriate contemporary groups has not been a problem as it has been for genetic evaluation of beef cattle.

Standardization. Several adjustments are applied to data before analysis. These adjustments reduce complexity of the evaluation model. Records are projected to 305 days and adjusted to a mature age, mean calving season, and standard genetic variance as well as for previous days open. Extremely low yields are raised to one-half the management group mean to minimize effect of these outliers (Norman and Dickinson 1989). An effect for age is included in the model to provide an additive adjustment in addition to the multiplicative adjustment for mature age. No adjustment is made for the use of bovine somatotropin because its use is not routinely reported for US herds; if its use were reported, cows that receive bovine somatotropin could be assigned to separate management groups.

Review of evaluations. Evaluations and supplemental information are reviewed. Most problems are identified by looking at maximums and minimums. As with evaluations of beef cattle, dairy bulls with large changes in evaluations are reviewed to determine if some systematic error has occurred with the data or the programs preparing output.

Estimates of genetic trend have been a useful indicator of the consistency of US yield evaluations. The age effect was added to the animal model because trend estimates from evaluations based on only first lactations were different from those from evaluations based on all lactations.

DISCUSSION

The goal of genetic evaluation is to provide accurate predictions of future progeny in support of selection decisions. Success depends on having both an appropriate model and adequate data. Field data are subject to many errors and require a sophisticated system of checks to ensure consistency and to eliminate errors. Because of breeder interest in individual animals, any elimination of data must be defendable.

An evaluation's value depends on the confidence that users have in its reliability. Although many data errors affect only individuals rather than the entire population, minimization of errors is critical to building trust in evaluations. A few individuals with unwarranted high evaluations can erode confidence in a genetic evaluation system. Producers buy, sell, and select individual animals, not populations.

Current US evaluation systems for beef and dairy cattle rely on cooperation of several organizations to supply data for computing evaluations. Continuing effort is underway to improve evaluation systems while coping with pressures from producers to reduce costs of data collection. Advances in technology provide some hope for reducing costs of data collection through automated animal identification and on-farm yield component determination. Reductions in cost of computing resources has enabled improvements in checking and availability of data and application of ever more realistic models.

REFERENCES

Christensen, L.G., Madsen, P. and Petersen, J. (1982) *2nd World Congress on Genetics Applied to Livestock Production* **7**:200-208.

Garrick, D.J., Pollak, E.J., Quaas, R.L. and Van Vleck, L.D. (1989) *J. Anim. Sci.* **67**:2515-2528.

Long, E., Johnson, R.K. and Keele, J.W. (1990) *J. Anim. Sci.* **68**:4069-4078.

Mallinckrodt, C.H., Golden, B.L. and Bourdon, R.M. (1992) *Proc. West. Sect. Am. Soc. Anim. Sci.* **43**:135-138.

Norman, H.D. and Dickinson, F.N. (1989) *J. Dairy Sci.* **72**:173-179.
Norman, H.D., Waite, L.G., Wiggans, G.R. and Walton, L.M. (1994) *J. Dairy Sci.* **77**:3198-3208.
Van Vleck, L.D. (1970) *J. Dairy Sci.* **53**:1697-1702.