ENHANCING QUALITY OF DYSTOCIA DATA BY INTEGRATION INTO A NATIONAL DAIRYCATTLE PRODUCTION DATABASE

C. P. Van Tassell[1,2] and G. R. Wiggans[1]

Animal Improvement Programs Laboratory[1] and Gene Evaluation and Mapping Laboratory[2]
Agricultural Research Service, United States Department of Agriculture
Beltsville, MD 20705-2350, USA

INTRODUCTION

The Animal Improvements Programs Laboratory (AIPL) of the United States Department of Agriculture (USDA) assumed responsibility for conducting the national genetic evaluation for dystocia (calving difficulty) and maintaining the associated database in 1999. The National Association of Animal Breeders (NAAB) supports the research, data collection and calculation of these evaluations.

To clarify terminology, all relationships discussed will be relative to the calf born to generate the dystocia observation. Specifically, the dam is the cow observed for calving difficulty, the sire is the service sire for this parturition, and the maternal grandsire (MGS) is the sire of the cow delivering the calf.

Interest in adding maternal effects to the genetic evaluation model arises because of concern about the antagonism between direct and maternal genetic effects on dystocia (Burfening, *et al.*, 1981; Thompson *et al.*, 1981; Manfredi *et al.*, 1991). However, only approximately half of the records submitted to AIPL include MGS identification (ID). This paucity of MGS ID is a problem when trying to model a maternal effect by using a sire-MGS model.

The dystocia data were migrated to a relational database that is integrated with the AIPL national database of production data originating from Dairy Herd Improvement Associations (DHIA) and includes lactations and pedigree back to 1960. This database was implemented to serve several purposes. First, a more rigorous series of data edits is possible by comparing with the production data (e.g., comparing calf birth dates in dystocia data with calving dates in the production data). Second, fast access to specific subsets of the data enables easier diagnosis and correction of problems. Finally, and most importantly, MGS ID rate could be increased by utilizing pedigree records in the production tables for records with dam ID sufficiently unique to allow matching with the production pedigrees.

MATERIALS AND METHODS

Data. An original master file was obtained from Dairy Records Management System (DRMS) when AIPL became responsible for conducting dystocia genetic evaluations. DRMS, a dairy records processing center, is contracted by NAAB to assemble and pre-edit dystocia records.

Subsequent update files are obtained twice annually. Data originate both from traditional DHIA data collection pathways as well as from AI organizations through collection by their cooperators.

Database design. Two primary tables were created for storage of dystocia data. One table contains the actual dystocia record. Most of the data fields currently recorded on the file format developed by NAAB are stored in this table. These fields include herd, sire, dam, calving difficulty score, parity of dam, calving date, multiple birth code, and data source. Sire and dam ID in this table are stored as animal keys, corresponding to the internal sequential numeric identifier assigned in the production data tables. The use of keys minimizes the problem of changed and multiple ID. The second table retains pedigree (i.e., MGS) information for records having dam ID that are not compatible with the production database. Fields in this table include dam and MGS ID, herd where the observation was recorded, and dam birth date.

When data are processed, the production database is queried for the dam ID from the dystocia record. If the dam ID is found, then that animal key is assigned. If a nonzero dam ID is present but not found in the production pedigree table and if MGS ID is present in the dystocia record, then a pedigree input record is generated and submitted to the production edit system to add the animal ID and pedigree information. If that pedigree record is successfully processed, animal key(s) are assigned for the dam and possibly for the MGS. MGS ID and birth date are updated if not present in the production pedigree, but existing data are not modified in the production database. For a record to be added to the AIPL dystocia database the sire ID must exist in the production tables because the purpose of this database is genetic evaluation of AI bulls. All bulls with assigned NAAB ID are included in the AIPL production database.

If the pedigree record is rejected by the edit system or the dam ID is zero or invalid, a negative key is issued for the dam. The use of negative keys facilitates the maintenance of the supplementary pedigree table for dystocia data because it avoids overlap with the production keys. A negative key in the calving ease table indicates that the pedigree information is stored in the dystocia table, while a positive key indicates that the information is in the production pedigree table. This scheme was designed to facilitate storage of pedigree information for the implementation of a sire-MGS model by allowing storage of MGS ID even if dam ID is missing or ambiguous. For nonzero dam ID both the production and dystocia databases are searched for the corresponding pedigree record before issuing a new key. Each record with unknown dam is assigned a unique negative key.

A number of data integrity edits are imposed, including breed of dam, appropriate values for dystocia, multiple birth codes and other required fields.

Detection of duplicate records. The data undergo preliminary editing to remove duplicates. Originally, based on NAAB suggestions for duplicate checking, records were considered duplicates if they contained the same herd, sire, calf birth date, parity of dam, and sex of calf. An additional constraint of equal dam is required if records originated from a single data source (i.e., dairy records processing center or AI organization). For records with positive dam keys identified, a more rigorous definition is applied: records with the same dam and calving dates within 6 months are considered duplicates. An additional class of duplicates is defined for records with negative

keys. Records with nonzero dam ID are considered duplicates if they have the same herd, the same dam key, and calving dates within 6 months.

The impact of the edits was evaluated by comparing the most recent master (RM) dystocia file with the database extract (DE) after data processing as described above. The same input files were processed by both systems. Only records for pure Holstein or Red & White breeding, for single birth calvings, and for births since 1980 were included in the analysis.

RESULTS AND DISCUSSION

Total number of records differed between RM and DE for a number of reasons. Not all the over 10 million records in RM were presented to the database, however, information on disposition of 8,861,363 records is available. Of these records, 105,414 were considered exact duplicates of existing records, i.e., all fields agreed between the records. Another 462,982 records were considered update records, where at least one data field differed between the records. A total of 33,378 records were rejected because of data problems - the most common reason was missing breed of dam. The remaining 8,259,589 records (>93%) were accepted. The DE contains 8,261,590 records which were drawn from all the accepted records and had multiple births, mixed breed and births prior to 1980 excluded.

The distributions of dystocia scores are shown in table 1. The data processing did not appreciably alter the distribution of scores in the total data set. Previous analyses of the data (Van Tassell and Sattler, 2000) demonstrated, however, that there is considerable variation in these distributions when evaluated on a herd or herd-year basis. Results from that study identified herds where scores clearly were not distributed like the population at large. With this knowledge, the extraction system was designed with the ability to review all a herd's data in order to decide to reject it if the frequency of extreme scores exceed arbitrary values. Additionally, MGS ID can be required and a minimum herd size can be imposed in data extraction.

Table 1. Distribution of dystocia scores for data from the recent master file and the database.

	Recent Master File		Database Extract	
Dystocia Score	Frequency	Percent	Frequency	Percent
1 - No Problem	7,701,559	76.55	6,267,362	75.86
2 - Slight Problem	1,041,203	10.35	883,228	10.69
3 - Needed Assistance	879,756	8.74	737,069	8.92
4 - Considerable Force	287,439	2.86	243,338	2.95
5 - Extreme Difficulty	150,875	1.50	130,593	1.58
Total	10,060,832		8,261,590	

As has been observed in previous studies (e.g., Berger, 1994), more difficult calving is observed for first parity than in later parities. Distributions of dystocia scores by parity for the DE data are

shown in table 2. Strong evidence exists for differences in calving ease between first and later parities, while the difference between second and later parities is relatively small.

Table 2. Distribution of dystocia scores for data from the database extract by parity with percentages within parity.

Dystocia	First Parity		Second Parity		Third and Later Parities	
Score[1]	Frequency	Percent	Frequency	Percent	Frequency	Percent
1	1,391,578	63.0	1,955,730	79.9	2,920,054	81.0
2	319,160	14.4	235,704	9.6	328,364	9.1
3	318,177	14.4	175,405	7.2	243,487	6.8
4	117,329	5.3	52,490	2.1	73,519	2.0
5	63,571	2.8	28,052	1.2	38,970	1.1
Total	2,209,815	26.8	2,447,381	29.6	3,604,394	43.6

[1]See Table 1 for definition of dystocia scores.

Percentage of male calves were 51.5 and 52.4 for RM and DE, respectively. Distributions of records across parities were also very similar for the two data sets, with 25.9 (26.8), 29.2 (29.6) and 44.9 (43.6) percent first, second and third or later parities for RM (DE) data.

The most important difference observed between the two data files was MGS ID rate. By integrating the data with the pedigree information, the rate of MGS ID increased from 57.2% (RM) to 73.1% (DE). Over 99% of the records with positive dam keys assigned had MGS ID recorded, while less than 13% of the negative key records contained MGS ID. A total of 69.7% of the records in the DE were assigned positive dam keys, indicating that over 30% of the dams being observed for dystocia are not uniquely identified.

CONCLUSIONS

By integrating pedigree information from production data, nearly 70% of the dam ID could be matched to the production pedigree table. The rate of MGS ID was increased from 57 to 73%, an improvement that will increase the accuracy of genetic evaluations when MGS effects are added to the evaluation model.

REFERENCES

Berger, P.J. (1994) *J. Dairy Sci.* **77**:1146-1153.
Burfening, P.J., Kress, D.D., and Friedrich, R.L. (1981) *J.Anim Sci.* **53**:1210-1216.
Manfredi, E., Ducrocq, V., and Foulley, J. L (1991). *J.Dairy Sci.* **74**:1715-1723.
Thompson, J.R., Freeman, A.E., and Berger, P. J. (1981) *J.Dairy Sci.* **64** :1603-1609.
Van Tassell, C. P. and Sattler, C.G. (2000) *J. Dairy Sci.* **83 (Suppl. 1)**:61.