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

G.R. Wiggans and L.L.M. Thornton
Animal Improvement Programs Laboratory
Agricultural Research Service, USDA, Beltsville, MD, USA
george.wiggans@ars.usda.gov

Data sources

- Dairy records processing centers - milk recording
- Breed registry societies - pedigree and conformation (type)
- National Association of Animal Breeders - calving traits and bull status
How data impacts accuracy

- Accuracy of recorded trait
  - Example: milk weight

- Emphasis and adjustment
  - Example: milking frequency, milkings weighed

- Other animals influenced
  - Example: parents, progeny, contemporaries

Pedigree and yield edits

- Identification (ID) verified for valid breed, country, and number
  - Canadian ID verified against Canadian Dairy Network data
  - Some American ID use last digit as internal check
Pedigree and yield edits (cont.)

- Birth date
  - Parent age checked (not too young and not too old for progeny)
  - Matched to dam calving date
    - Differences of <1 month allowed
    - Omitted if embryo-transfer animal

Pedigree and yield edits (cont.)

- Birth date (cont.)
  - Parents not previously in database added with estimated birth date
    - 3 years before reported animal’s birth date
    - Revised as data from older siblings received
Pedigree and yield edits (cont.)

- **Alias detection**
  - Same birth date and full siblings but not twins
  - Within-herd ID (control number) useful in identifying additional ID
  - Bulls registered in >1 country common cause

- **Alias detection (cont.)**
  - Numbers differing by single digit investigated as possible invalid ID
  - Yield data must not conflict for data from 2 ID to be combined as data for the same cow
Pedigree and yield edits (cont.)

**Yield**

- Values outside widest range rejected
- Values outside more narrow range stored but changed to a floor or ceiling if used
- Cow test date checked against herd test date

**Calving date**

- Cannot overlap previous lactation
- Missing calving date may cause breeding to be associated with previous calving
Error records

- Errors and conflicts stored in a record and returned to processing center to assist in data correction
  - **Reject** - record rejected
  - **Notify** - input record accepted but a problem may exist
  - **Change** - input record changed to match master

Error records (cont.)

- Stored to assist in answering queries
- Sometimes forwarded by processing center to milk-recording supervisor or producer for action
- Rejected records also available by query on web site - [http://aipl.arsusda.gov](http://aipl.arsusda.gov)
## Error frequency for pedigree records*

<table>
<thead>
<tr>
<th>Error</th>
<th>Simple definition</th>
<th>Action</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Nd</td>
<td>Merging input to animal in master</td>
<td>Notify</td>
<td>207</td>
</tr>
<tr>
<td>10h</td>
<td>Update input to twin</td>
<td>Change</td>
<td>138</td>
</tr>
<tr>
<td>3lb</td>
<td>Dam ID differs from master, source not verified</td>
<td>Notify</td>
<td>107</td>
</tr>
<tr>
<td>10d</td>
<td>Sibling updated to twin</td>
<td>Notify</td>
<td>106</td>
</tr>
<tr>
<td>2Be</td>
<td>Sire ID not preferred</td>
<td>Change</td>
<td>82</td>
</tr>
<tr>
<td>3Be</td>
<td>Dam ID not preferred</td>
<td>Change</td>
<td>75</td>
</tr>
<tr>
<td>5Fc</td>
<td>Birth date and dam calving date not the same</td>
<td>Notify</td>
<td>69</td>
</tr>
<tr>
<td>2jc</td>
<td>Sire ID differs from service sire ID</td>
<td>Notify</td>
<td>64</td>
</tr>
<tr>
<td>2lb</td>
<td>Sire ID differs from master, source not verified</td>
<td>Notify</td>
<td>60</td>
</tr>
<tr>
<td>4jc</td>
<td>Master same as cross-reference</td>
<td>Change</td>
<td>52</td>
</tr>
</tbody>
</table>

*n = 12,000

## Error frequency for lactation records*

<table>
<thead>
<tr>
<th>Error</th>
<th>Simple definition</th>
<th>Action</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>2De</td>
<td>Grade sire misidentified</td>
<td>Reject/</td>
<td>2,738</td>
</tr>
<tr>
<td></td>
<td></td>
<td>change</td>
<td></td>
</tr>
<tr>
<td>1Be</td>
<td>ID not preferred ID</td>
<td>Change</td>
<td>2,611</td>
</tr>
<tr>
<td>6Td</td>
<td>Parity and age mismatched</td>
<td>Change</td>
<td>2,334</td>
</tr>
<tr>
<td>0jd</td>
<td>Multiple birth code ignored</td>
<td>Change</td>
<td>2,235</td>
</tr>
<tr>
<td>7lc</td>
<td>Abnormal recorded milk yield</td>
<td>Change</td>
<td>1,967</td>
</tr>
<tr>
<td>2Gd</td>
<td>Sire ID differs from master</td>
<td>Change</td>
<td>1,902</td>
</tr>
<tr>
<td>7Ob</td>
<td>Quality control code incorrect</td>
<td>Notify</td>
<td>1,899</td>
</tr>
<tr>
<td>5bd</td>
<td>Birth date differs from master</td>
<td>Reject</td>
<td>1,801</td>
</tr>
<tr>
<td>3Gd</td>
<td>Dam ID differs from master</td>
<td>Change</td>
<td>1,707</td>
</tr>
<tr>
<td>7Mb</td>
<td>Milking weighed not the same as for herd</td>
<td>Change</td>
<td>1,472</td>
</tr>
</tbody>
</table>

*n = 93,000
Error records query

Editing principles

- Data either rejected or modified when errors encountered
- Effect of rejection
  - Loss of possibly valuable information
  - No genetic evaluation for animals of interest
- System designed to retain data whenever possible
- Data elimination preferred to retention of conflicting data
Example

- Animal’s birth date conflicts with dam’s calving date
- Both animals already have data in system
- Dam ID removed to resolve conflict and to allow records for both animals to remain in database

Importance of types of data

- Milking times
- Alternation of supervised milking
- Herdmate identification
- Breed reporting for crossbreds
- Data collection rating
- Automatic milk recording
Milking times

- Most herds enrolled in a.m.-p.m. testing
  - Not all milkings supervised
  - Daily yield estimated from recorded milking based on interval since previous milking
- Start and end times required because of variation in length of milkings
- Most accurate estimate of interval between milkings derived from midpoints of consecutive milkings instead of start times

Alternation of supervised milkings

- National formulas to estimate a.m.-p.m. yield not an exact fit for individual dairies
- Alternation of supervised milkings between morning and evening
  - Averages out systematic errors over time
  - Difficult to achieve with large herds
Herdmate identification

- Genetic evaluations rely heavily on pedigree data
- Data from cows with unknown sires not included in evaluations
- Only evaluated cows used as herdmates for other cows
- Large herds may have small contemporary groups if most cows not sire identified

Crossbred breed reporting

- U.S. genetic evaluation is across breeds
- Breed percentages derived from pedigree
- Breed determines breed base on which cow’s evaluation is reported unless breed coded as XX (crossbred)
- Sire breed determines breed base for evaluations of crossbred cows
Crossbred breed reporting (cont.)

- Animal’s breed should reflect breed with highest percentage from within animal’s pedigree
- Genetic evaluations for crossbred herds likely to be reported on different breed bases
- For animals with equal breed percentages, using predominant breed for herd is beneficial

Data collection rating (DCR)

- Measures how much information was collected relative to a standard test plan
- The less information collected, the lower the DCR and the higher the error variance
- Does not measure bias directly
DCR (cont.)

- **Example:** Same milking sampled every month under a.m.-p.m. testing with component sampling
  - Component estimates biased by degree that national estimation formulas do not fit herd
  - Amount of information collected not different
  - Error variance not increased
  - DCR the same

DCR (cont.)

- DCR for unsupervised milkings arbitrarily set to 75% of that for a supervised milking
- Similarly discounted DCR could be used for herds enrolled in a.m.-p.m. testing
Automatic milk recording

- Opportunity for increased recording accuracy
  - Must monitor own accuracy and detect when unit needs maintenance
  - Dependent on accurate cow ID
- 5- to 10-day averages usually reported
- Atypical cow yields detected and excluded
- Accurate meter calibration important

Conclusions (cont.)

- Highly complex system for checking data used in national U.S. genetic evaluations of dairy cattle
- Conflicting data from various sources
  - Harmonized based on which data are expected to be most accurate
  - Deleted when necessary
Conclusions

- Evaluation accuracy dependent on accuracy of all contributing data
- Invalid records diminish evaluation accuracy of evaluations for other animals

Thank you

- John Clay
  Dairy Records Management Systems
- Dan Webb
  University of Florida
- Lillian Bacheller
  AIPL, ARS, USDA