Embracing new streams of dairy data - how good is good enough?

J. Mattison and S. Sievert

National DHIA/Quality Certification Services, P.O. Box 930399, 53593, Verona, WI, USA
Corresponding Authors: jmattison@dhia.org; sjsievert@dhia.org

Abstract

With the investment of new technologies for on-farm data measurement and recording, there is an increasing amount of data available from dairy operations. These data include estimates of traditional measures such as milk yield and components, but also provides access to measurements on new traits of interest to herd recording organizations. With the volume of data available, it is logical to find new paths for data movement from on-farm software or manufacturer servers into organization databases. ICAR has supported milk recording organizations and dairy producers in the traditional test-day collection of data with standards and equations that deliver data for daily management decisions and to qualified information for genetic evaluations. Today, the challenge that exists for recording organizations and national databases is balancing the volume of data from new sources with the accuracy of the individual measurements. Further, the speed of data flow has the potential to allow unqualified data to enter databases before issues or concerns are identified. The challenges of veracity, volume and velocity of new streams of data lead to the question 'how good is good enough' for usability by organizations and databases.

In addition to identification of the source, an evaluation of the data quality at a system level is key for determining the usability or suitability. Focus on the automatic identification as well as the automatic recording of measurements is critical, both on-farm and as data moves into central databases. In addition to linkage of animal ID to measurements, considerations on the completeness and consistency of continuous data flow are of merit. Finally, the editing of data points, including estimation of missing observations and the precision of individual measurements are a component of data quality and contribute to the usability of data. A fresh approach on the inclusion of new streams of data is needed. This systems quality approach should embrace the technology of on-farm sensors that generate valuable insights into dairy cattle milk yield, milk composition, conformation, behavior and activity with the goal of inclusion of usable data that is 'good enough' rather than focus on data exclusion based on precision of an individual measurement.

Background and perspective

Herd recording programs have traditionally relied on data collected on the dairy during periodic visits and includes milk yields and cow event data accompanied by visible cow identification by organizational staff. The source of the yield data, and subsequent milk component estimates, is ICAR-certified devices either owned by the herd recording organization or by the dairy operation. Even when properly installed and operated under ideal circumstances, the data provided is limited to the recording day and only provides an estimate of cow performance during that timeframe. This system has provided excellent data for both management systems and genetic evaluations as the

accuracy of these measurements is essential for modelling the performance of cows across the interval between visits. However, the development of new technologies for measurement of various cow traits that include both traditional yield and component measures, but also data that physically describes cow conformation, movement, and even provides estimates of cow health and well-being. The adoption of these technologies by dairies has led to more data available to herd recording organizations and their customers cooperatively providing data with the expectation of real-time access to results and information for management decisions affecting all aspects of the dairy operation.

While there is, and will be, a segment of dairy operations participating in herd recording programs in the traditional fashion, there are opportunities for new streams of dairy data that may be used and useful to herd recording organizations. However the question that faces the industry with these data streams is how or can we use these data? The volume of data and the velocity of data generation is staggering relative to traditional data ingestion models. But with this data flow also comes an increased variation in measurements and questions on the validity of the data as it flows automatically from the dairy operation to external databases. In summary, how good is good enough? This paper does not propose to set standards or guidelines for use of data but does offer a review of considerations for herd recording organizations as they look to development of programs and services offered to the modern dairy operation.

Data use and quality

New streams of dairy data should be reviewed for the use, useful and usability of the data on various levels. Data used for management decisions or animal health/welfare metrics will most likely have different measures of quality associated with it compared to data used for genetic evaluations. However, this does not mean that data use is exclusionary for these purposes, rather there needs to be an evaluation of data sources and continuous monitoring of quality. One of the key considerations with new streams of dairy data is the dynamic nature of data collection/measurement. Many of these systems rely on the simultaneous recording of animal identification and measurement of the traits(s) by the system. While the location or pen identification may be more important for certain management traits relating to feeding or milking metrics, the linkage of animal identification used for data measurement with the official identification of the animal is an important part of the data quality. Herd recording organizations should have or will need to develop protocols for evaluating both on-farm and data transfer for identification linkage.

As previously noted, data quality needs to be monitored as we look to new streams of dairy data. Traditionally, the accuracy of the individual measurement has been considered critical to the assessment of usability and said accuracy is an important part of data quality. However, quality data has additional elements that must be considered as we look to use and usability in the volume of data available on many dairy operations. These quality elements include accuracy, completeness, consistency, credibility and custody of the data. A continuous approach to evaluation and monitoring of data quality is paramount for herd recording organizations. For some data measurements, bias and precision are important are varying levels, but other sources of error may negatively affect the usability of data for that metric or measurement.

While some of the quality elements can be identified and managed by assurance programs, there are elements that require external review by ICAR or other organizations to understand data handling in both farm management software and as the data moves to external databases. These include an understanding of data smoothing, estimation of missing measurements, and the distribution of errors relative to the collection system. External validation of these systems along with internal assurance programs

is essential for determining usability however should not be viewed as exclusionary for data ingestion by herd recording organizations. Rather, consistently recording the source of the data measurement(s) in the data flow will allow for usability decisions to be made as the use of data is determined.

How good is good enough?

There is no simple answer to the question 'how good is good enough' in this presentation or in practice. However there is opportunity for new and dynamic approaches to evaluate new streams of dairy data by herd recording organizations. This approach should have five key considerations that include:

1. Validation of system data quality and recording the source of data measurements in the data flow.

2. Focus on the ingestion of data from farm management software instead of exclusion of data based on measurement accuracy.

3. Development of routine data quality monitoring tools to assure consistent and credible data.

4. Focus on the data use, usefulness, and usability for specific needs rather that an 'all or nothing' approach to data use.

5. Efficient removal of incomplete data instead of attempts to edit or standardize the data to meet arbitrary standards or improve the data quality.

This approach is a departure from traditional data collection and use schemes but is essential with new streams of dairy data. The volume and velocity of data may be viewed as a deterrent but with a data quality monitoring program, there is opportunity to deliver information to dairy operations, herd recording organizations, and national/multi-national databases.