Proc. Assoc. Advmt. Anim. Breed. Genet: 24: 139-142

CHARACTERISING THE QUANTITY AND QUALITY OF DATA USED IN MERINO SHEEP GENETIC EVALUATION SYSTEMS

S.Z.Y. Guy[1] and D.J. Brown[1]

[1] Animal Genetics Breeding Unit*, University of New England, Armidale, NSW, 2350 Australia

SUMMARY

Estimated Breeding Values (EBVs) published by Sheep Genetics Australia have an accuracy estimated with them. While the EBVs, their accuracy, and errors of genetic parameter estimates are all influenced by both data quantity and quality, these calculations do not explicitly take into account all aspects of data quality. To encourage increased genetic gains, Sheep Genetics provides participating breeders with data quantity and quality metrics in a 'RAMping Up Genetic gains' report. This paper demonstrates the considerable variation in these metrics for Merino flocks, and proposes additional descriptors metrics to characterise the quantity and quality of sheep genetic evaluation data. Current results show that there are opportunities to improve the completeness of pedigree and reproduction trait recording. Flocks had on average 46.6 ± 36.1% (mean ± SD) of animals with full pedigree, and 4.1 ± 6.9% of animals within each flock with reproduction trait records. The average proportion of effective progeny was 64.3 ± 19.1%. Flocks had on average 40.2 ± 37.3% of animals in contemporary groups that had variation in birth date recording. Since variation in age within contemporary groups is expected, this highlights potential issues with accurate recording of birth dates. Additional metrics describing lambing date distributions and deviations from the expected dates were derived, and emphasise potential issues of birth date accuracy, with some flocks recording birth dates on a non-random proportion of days of the week. Feedback on the quantity and quality of their current data should help ram breeders target improvements on their recording program. However, the optimum or reasonable level of quantity and quality to maximise genetic gains is currently undefined.

INTRODUCTION

The genetic evaluation systems available to the Australian sheep and beef industry through Sheep Genetics and BREEDPLAN, respectively, primarily rely on industry data submitted by seedstock producers. While there are standards and guidelines, there is wide variation in the data submitted. An accuracy figure is reported alongside estimated breeding values (EBVs). While the quality of data has been shown to influence the EBVs, their accuracy and the errors of genetic parameter estimates, accuracy is calculated using the amount and structure of information utilised (i.e. quantity), and not explicitly the quality of information. The difference between data quantity and quality is highlighted in the following example; a date of birth may be supplied for each animal (maximum data quantity), but a single generic date may be used for all animals irrespective of their actual date of birth within the lambing period (poor data quality). This will affect the ability to accurately correct for age and thus the accuracy of the EBVs. This highlights the need for additional metrics beyond EBV accuracy to characterise the quality of data.

Data Quality Grades, which reflect the level of recording for pedigree, scan and wool traits, were previously provided to LAMBPLAN clients as a practical approach to describing index accuracy (Banks, 1999). Currently, Sheep Genetics provides the 'RAMping Up Genetic gains' (RUGG) report to participating breeders, which includes metrics to describe the quantity and quality of pedigree and

* A joint venture of NSW Department of Primary Industries and the University of New England

performance recording, and data structure. These metrics have been shown to have an association with genetic gains for a flock (Stephen *et al*. 2018). This paper demonstrates the variation in the data quantity and quality metrics reported in RUGG reports for Merino flocks, and proposes additional metrics to characterise the quantity and quality of data being supplied to Sheep Genetics.

MATERIALS AND METHODS

Existing data metrics. The metrics reported in RUGG reports were available for the 265 Merino flocks from the 12th December 2020 analysis. These flocks had a minimum of 100 animals per year and data available for the last 5 years. Unless stated, metrics were calculated as an average of the last 5 years and across contemporary groups. Metrics were classified as either quantity or quality metrics, although it must be acknowledged that some metrics can be placed in either category:

1) **Quantity:** the amount of data submitted and its completeness.
- **fullped** (%): proportion of animals from the flock in the analyses where both sire and dam are known (i.e. full pedigree).
- **avpedknown** (%): completeness of pedigree known from last 3 generations.
- **recorded** (%): proportion of animals with records submitted for any of the following: weight, fat, eye muscle depth, fleece weight and fibre diameter (all age stages) or number lambs weaned.
- **ngeno** (%): proportion of animals genotyped.

2) **Quality:** the appropriateness for its intended use, including accuracy and data structure.
- **synped** (%): proportion of animals with syndicate pedigree (i.e. where multiple rams are mated over a group of ewes, resulting in multiple potential parents for the progeny).
- **ages** (%): proportion of animals recorded that are in contemporary groups with variation in age. Variation in age within contemporary groups is expected with accurate birth date recording.
- **bt** (%): proportion of animals recorded that are in contemporary groups with variation in birth type recorded.
- **eff** (%): proportion of effective progeny (i.e. percentage of progeny from a given sire relative to all progeny within a group, as defined in Brown *et al.* 2001).

Additional quantity metrics. To take into account the different breeding objectives of each breed type, a 'recorded' metric was expanded to the proportion of animals recorded by trait groups:
- **rec_weights**: weight traits, ultrasound c-site fat depth, and ultrasound eye muscle depth.
- **rec_repro**: number of lambs weaned.
- **rec_wool**: greasy fleece weight and fibre diameter.

Additional qualtity metrics. These included genetic linkage metrics by trait group, as well as metrics to describe lambing date distributions and deviation from uniform distributions (inspired by DataAudit and StockTake; Johnston and Moore, 2005):
- Average proportion of animals recorded that are directly linked to external flocks, by trait groups: carcase scan traits (**link_carcase**), weight traits (**link_weights**), number lambs weaned (**link_repro**), wool traits (**link_wool**).
- **maxfreq_ywt** (%): the percentage of the most common single value appearing. Missing values were not included in this calculation. Only results for yearling weight (ywt) are reported in this paper as ywt was the most common weight trait recorded for the flocks examined.
- Chi-squared statistics**:** For a large sample size of data with sufficiently wide variation in values, the last digits are expected to have a uniform distribution (Dlugosz and Müller-Funk, 2009). Deviation from this expectation may be due to poor recording, equipment problems or non-randomisation of recordings. Since different traits are recorded in various increments (e.g. as whole number integers or various decimal places), chi-squared statistics were calculated for the last digits in the units (**chi_units_ywt**) and tenths (**chi_tenths_ywt**) place values for ywt:

$$\chi^2 = \sum_{i=0}^{10} \frac{(\text{expected\%} - \text{observed\%})^2}{\text{expected\%}}$$

where expected = 10%, and observed = % of records with the digit *i*

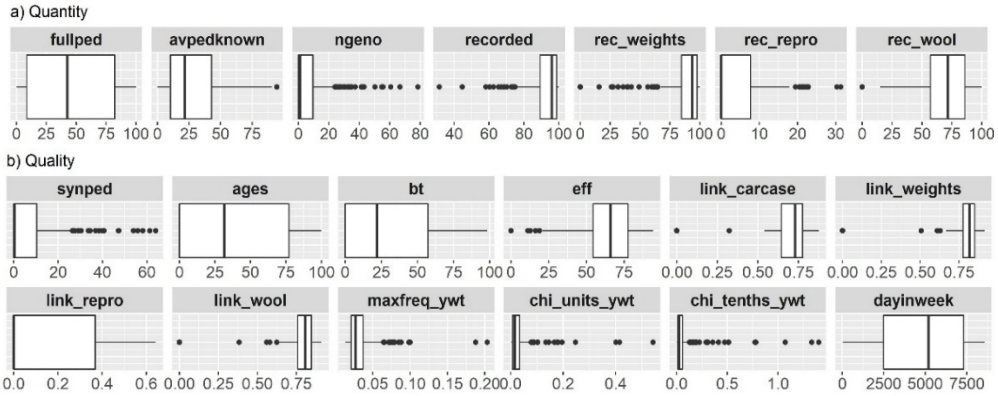
- **dayinweek**: mean square error of birth date for days in week. This metric is based on the same concept as the chi-squared statistics, where the likelihood of birth dates to occur on any given day of the work is expected to be equal. This was calculated as:

$$\text{dayinweek} = \sum_{i=1}^{7} (\text{expected\%-observed\%})^2$$

where expected = 1/7 × 100 % for each day of the week, and observed= % of animals born on the *i*th day of the week.

RESULTS AND DISCUSSION

Variation in data metrics. Figures 1a and 1b demonstrate the considerable variation in the data quantity and data quality metrics across Merino flocks.



The data quantity metrics describe the amounts of pedigree and performance recording across Merino flocks. The average proportion of animals within a flock with full pedigree (*fullped*) was 46.6% ± 36.1% (mean ± SD), with 29.2± 36.1% of the pedigree complete over the last 3 generations (*avpedknown*). Flocks had an average of 7.4% of animals within the drop genotyped (*ngeno*). Recording by trait groups was more informative than a recording metric that included all traits. As expected, there was more recording for weight traits (*rec_weights*, 86.9 ± 19.2%) and wool traits (*rec_wool*, 69.0 ± 20.7%) than reproduction traits (*rec_repro*, 4.1 ± 6.9%). Since only a proportion of ewes enter the ewe flock, low values of *rec_repro* were as expected. However, it was also the most variable metric relative to the mean (range 0 to 31.3%, CV= 168.0%). These metrics highlight the opportunity for Merino breeders to improve recording for pedigree and reproduction traits.

The data quality metrics describe varying levels of pedigree accuracy and distribution of data. There was a low proportion of animals with syndicate pedigree (*synped*, 7.4 ± 12.3%). However this metric was also the most variable (range 0 to 63.9%, CV = 164.9%). The proportion of animals in a contemporary group that had variation in recording for birth dates (*ages*) was 40.2 ± 37.3% and 32.0 ± 32.1% for birth type recording (*bt*). That is, ~60% animals were in groups where there was no variation in birth date, and ~68% with no variation in birth type. This highlights potential issues with accurate recording of birth dates and birth types. The average proportion of effective progeny (eff) was 64.3 ± 19.1%. Since the eff metric can only be estimated if sire pedigree is known, this is expected to be an underestimate. The degree of linkage to other flocks reflected the level of recording

by trait group and Merino breeding objectives, with the most linkage through weight, wool and carcase traits compared to reproduction traits (*link_weights*, 78.6 ± 15.8%; *link_wool*, 78.1 ± 13.6%; *link_carcase*, 58.2 ± 31.3%; *link_repro*, 14.3 ± 20.1%). An average of 3.3% of yearling weight records (*maxfreq_ywt*) were the same within each flock (range of 0.01% to 20.3%).

The quality metrics describing distributions of traits and deviations from expected distribution also varied across flocks. The chi-squared statistics, describing last digit distributions, were all less than the chi-squared critical value of 3.325, suggesting that the frequencies of last digits for ywt were as expected. Conversely, the average *dayinweek* was 4,841.5 ± 2777.53, and ranged from 20.3 to 8,571.0 (i.e. the maximum mean square error, with birth dates recorded on only one day of the week). Again, the required degree of accuracy for birth dates and what is considered a reasonable loss in age adjustment precision is currently unknown. Nevertheless, these distribution and deviation metrics can still be used as a way to highlight unusual data.

Relationships between metrics. The relationships between the quantity and quality metrics were quantified by Pearson's correlations (*r*). As expected, there were strong linear associations between *rec_repro* and *link_repro* (*r* = 0.82), and *fullped* and *avpedknown* (*r* = 0.77). There were moderately strong associations between *fullped* and *ages* (*r* = 0.60), *bt* (*r* = 0.61), *daysinweek* (*r* = -0.51), *link_repro* (*r* = 0.48) and *eff* (*r* = 0.40). There were also strong associations within categories (e.g. between *ages* and *daysinweek*, *r* = -0.85). Therefore, the quantity and quality metrics are not necessarily independent, and some metrics describe similar aspects.

Industry implementation. The improvement of the quality and quantity of data, in particular for reproduction traits, has been identified as a key priority for Sheep Genetics (Collison *et al.,* 2018). A framework to characterise genetic evaluation data, including a carefully developed overall 'data quality score, will benefit individual breeders, ram buyers and the industry as a whole. Feedback on the quantity and quality of their current data will allow ram breeders target improvements on their recording program, which support selection decisions and maximise genetic gains, and assess changes in recording across time. A data quality score could also help identify and highlight breeders who collect high quality data. In turn, this will provide increased transparency to ram buyers about the quality of data used to calculate EBVs. There is also potential to use these metrics to determine how data contributing to the reference population is valued and rewarded.

CONCLUSIONS

This paper demonstrates the considerable variation in the quantity and quality of Merino sheep genetic evaluation data. While there are opportunities for Merino flocks to improve completeness and accuracy of pedigree recording, birth date and reproductive performance, the optimum or reasonable level of quantity and quality is currently undefined.

ACKNOWLEDGEMENTS

This work was funded by MLA project L.GEN.2004. The authors also grateful for the AGBU and Sheep Genetics staff who provided valuable feedback and discussion on this topic.

REFERENCES

Banks R.G. (1999) *ICAR Technical Series.* **3**: 521.

Brown D.J., Tier B. and Banks R.G. (2001) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **14**: 441.

Collison C.E., Brown D.J., Gill J.S., Chandler H.R., Apps R., Swan A.A. and Banks R.G. (2018) *Proc World Congreess on Genetics Applied to Livestock Production*. **11**: 661.

Dlugosz S. and Müller-Funk U. (2009*). Adv. Data Anal. Classif*. **3**: 281.

Johnston D.J. and Moore K.L. (2005) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **16**: 161.

Stephen L.M., Brown D.J., Jones C.E. and Collison C.E. (2018) *Proc World Congreess on Genetics Applied to Livestock Production*. **11**: 433.