The Data Quality Score: objective assessment of data quality for Australian sheep breeders

D.J. Brown[1*], E.J. McCrabb[2], P.E. Bradley[2], I.J. Rose[2], R.G. Banks[1] and S.Z.Y. Guy[1]

[1] AGBU, a joint venture of NSW Department of Primary Industries and University of New England, 2351, Armidale, Australia; [2] Meat and Livestock Australia, The Short Run, University of New England, 2351, Armidale; [*] dbrown2@une.edu.au

Abstract

Data quality influences the accuracy of estimated breeding values, and hence the accuracy of selection and genetic progress. This paper describes how a data quality score (DQS) was created to characterise the overall data quality for Australian sheep flocks. The data quality metrics investigated captured data quantity (e.g. degree of performance/pedigree recording), accuracy, structure (e.g. sire representation and linkage) and timeliness of data submission. These metrics were combined to calculate the overall DQS, with weightings dependant on breed type, variation and relationships between metrics. The DQS was well received by industry, with overwhelming support to publish the DQS after an initial producer education campaign. The DQS will help identify and highlight breeders who collect high quality data, help breeders improve their data quality, and increase information available to ram buyers. This research supports Australia's sheep industry to value phenotypes in an objective manner.

Introduction

Data is the cornerstone of genetic evaluation systems. Maximising data quality enhances accuracy of estimated breeding values and of selection decisions, which drives higher rates of genetic progress. The level and breadth of recording, and the availability and accuracy of supporting information (pedigree, fixed effects and genotypes), also impact selection accuracy and genetic progress. The genetic evaluation for the Australian sheep industry, Sheep Genetics, primarily relies on data submitted by ram breeders. While there are standards and guidelines for data submission, there is wide variation in data recording and submission across the industry.

Sheep Genetics make available 'RAMping Up Genetic Gains' (RUGG) reports, which provide breeders with feedback on their data (Stephen *et al.*, 2018). While these RUGG reports are available to breeders and have been used as an engagement tool for service providers and consultants, uptake by industry has been varied. As an incentive to seedstock breeders to improved recording, a data quality score (DQS) was developed to describe (and potentially rank) the quality of data submitted for genetic evaluation. This paper describes metrics that characterise data quality, the development of the DQS, and the response to the DQS by Australian sheep breeders.

Materials & Methods

Data quality metrics. The following data quality metrics were calculated for each flock. Full Pedigree (%): the overall proportion of animals from the flock in the analyses with full pedigree. Contemporary group variation in 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. Proportion of effective progeny (%): calculated as the proportion of progeny each sire has in the group, averaged across sires and all contemporary groups. Proportion of animals recorded for the weight, reproduction, wool and carcase trait groups (%). Average proportion of animals recorded that are directly linked to external flocks, by trait group (%): weight, carcase scan, reproduction and wool. Timeliness of data submission was calculated as interval between the date of recording and actual submission date (days). Calculation of the data quality metrics that capture the above factors are provided in more detail in Guy and Brown (2021a). These metrics were calculated for 441 Terminal, 304 Merino, and 96 Maternal flocks.

The Data Quality Score. The metrics above were combined into an overall DQS for each flock. All metrics were weighted considering breed type (Terminal, Maternal or Merino), variation and relationships within the data quality metrics and relationships between data quality metrics and metrics describing genetic gains (Guy and Brown, 2021b). Due to the different scales and variances of each data characteristic metric, they were scaled to unit variance and a mean of zero. The resulting DQS for each method was also then scaled to between 0 and 100 for ease of interpretation.

Consultation with industry. The reporting of the data quality metrics and DQS was evaluated at 6 industry events involving 96 flocks. DQS reports were generated and incorporated into the existing RUGG reports. Reports were provided for each participating flock, followed by group discussion and collection of anonymous feedback.

Results

Data quality metrics. The data quality metrics with notable variation are shown in Figure 1.

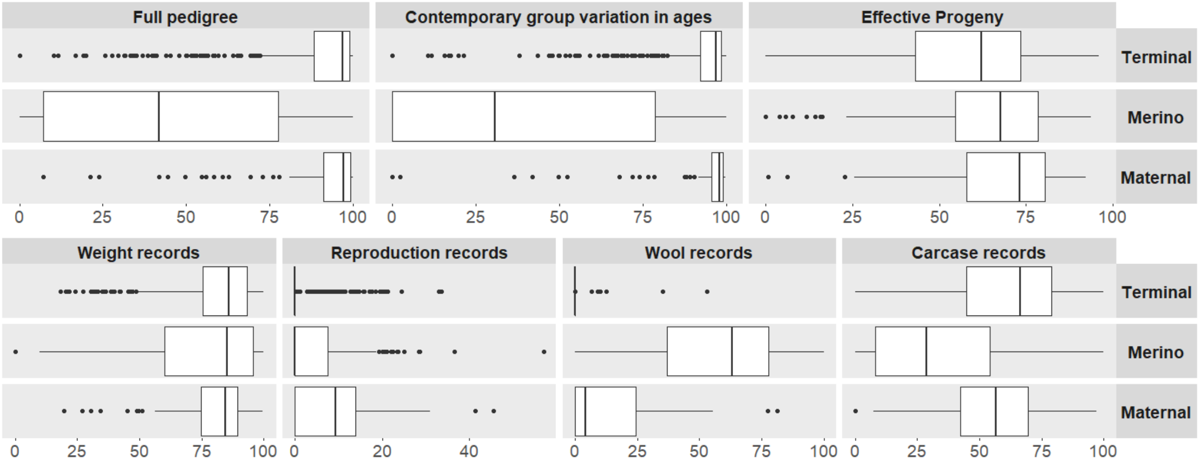

Figure 1. Distributions of data quality metrics for Terminal, Merino and Maternal flocks.

While distributions for different breed types overlapped, Merino flocks had markedly lower full pedigree recording and more variation across flocks (mean ± SD, 43.6 ± 35.3%) compared to Maternal (89.3 ± 18.6%) and Terminal (87.2 ± 21.6%). There was also less variation in ages within Merino contemporary groups. Given that the majority of lambs originate from natural mating this implies less accurate recording of birth dates.

The data quality metrics relating to data structure were fairly consistent across breeds. Average effective progeny numbers were 67.5 ± 18.2% for Merino, 64.0 ± 20.4% for Maternal, and 57.5 ± 21.1% for Terminal flocks. The level of linkage for weight traits was 76.0 ± 19.7% in Merino, 68.1 ± 13.4% in Terminal and 64.5 ± 16.9% in Maternal flocks.

All breed types had a high proportion of animals recorded for weight traits, with Merino flocks having the lowest level of recording, and the most variability across flocks within the breed type (75.2 ± 25.1% compared to 79.8 ± 15.9% for Maternal and 81.4 ± 17.2% for Terminal flocks). Maternal flocks recorded more reproduction traits and had the most variation compared to other breed types (8.7 ± 8.9%, compared to 4.2 ± 7.6% for Merino and 1.8 ± 4.9% for Terminal flocks). Merino flocks had a greater proportion of animals with wool traits recorded, although this varied greatly (59.3 ± 25.9%). Terminal flocks had the highest level of carcase phenotype recording (61.4 ± 20.1%, compared to 54.2 ± 20.1% for Maternal and 33.2 ± 28.4% in Merino flocks).

There was also variation in how timely data were submitted for analysis. Merino flocks took longer to submit pedigree records (average age of animals to appear in the pedigree was 376 days, compared to 240 days in Terminal and 243 days in Maternal flocks). The time taken to submit phenotypic data was similar across breeds for weight traits (submission time ranging from 92 days to 109 days post-measurement) and carcase data (range 38 days to 55 days).

The Data Quality Score. Metrics relevant to each breed type were included in the calculation of the DQS (for example, level of recording and linkage for wool traits were not included for the DQS calculations for Terminal flocks). Specific metrics were also given higher weights to target improvements for each breed type (for example, degree of full pedigree recording was lowest for Merino flocks, so more emphasis was placed on this metric for breeders in that group).

The DQS was categorised as 'star ratings' by equally splitting the DQS range (score of 0-20: 1 star, 20-40: 2 stars, 40-60: 3 stars, 60-80: 4 stars, 80-100: 5 stars; Figure 2).

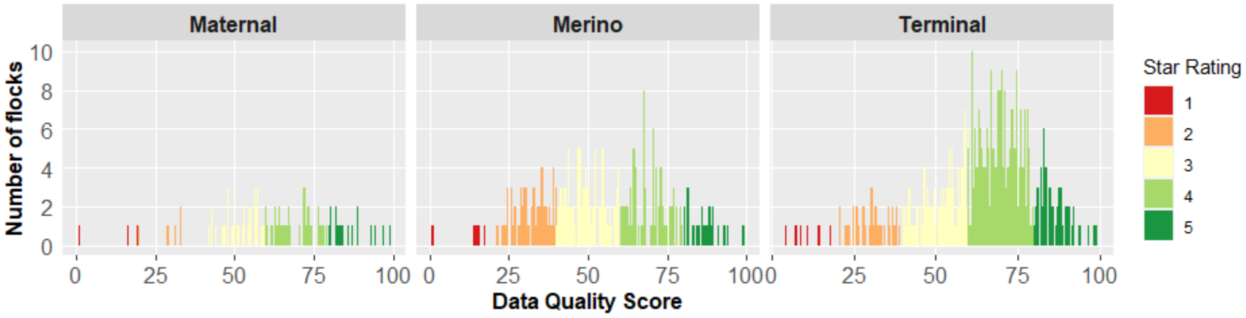

Figure 2. Distributions of Data Quality Score for Terminal, Merino and Maternal flocks.

Consultation with industry. The DQS was well-received by industry. Constructive feedback was provided to further enhance the usefulness of the DQS and the reports developed to support it. The majority of anonymous survey respondents thought the DQS and star rating were useful and relevant to industry, and also a source of motivation to improve their own recording strategies (Figure 3).

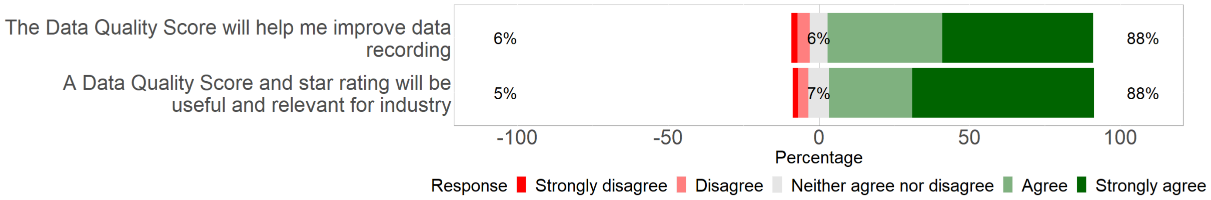

Figure 3. Level of agreement (green) and disgreement (red) for the usefulness of the Data Quality Score for improved recording, and relevance for industry (63 responses).

There was also overwhelming support to report the DQS to breeders, and to publish the data quality star ratings for each flock following an extension effort to help ram breeders and buyers understand the new information.

Discussion

Data quality metrics. The level of performance recording for each trait aligned with the general breeding objectives of each breed type. The high proportion of weight trait recordings across all breeds is likely to reflect the relative ease of recording compared to other traits. There was significant variation across breeds and flocks for all other trait groups, which highlights the opportunity for improvement in all breed types.

Delivering genetic progress through data quality. Data quality metrics were related to metrics describing genetic gains (average index accuracy, average index value and index trend) in Merino flocks (Guy and Brown, 2021b), as well as Maternal and Terminal flocks (unpublished). This suggests that improvement in data quality should lead to improvement in the rates of genetic progress.

Some aspects of data quality are captured in estimated breeding value (EBV) accuracies, which are reported alongside EBVs. While the accuracy of genetic merit estimation is influenced by the level of recording and management group structure (Brown *et al*. 2001; Swan and Brown, 2007), EBV accuracy is calculated using the amount and structure of information utilised (i.e., quantity), and not explicitly the precision and timeliness of information supplied to the analysis. The reporting of additional data quality metrics beyond EBV accuracy provides breeders with targeted feedback to help them improve their data recording programs.

In addition to the general reporting of these new metrics to industry, more extension and decision support tools are required to help breeders implement sensible performance recording improvements. Further work planned is to conduct cost-benefit analysis of different recording options and embed this into tools which can be used by service providers as the basis for objective breeding program advice.

Benefits to industry. The DQS will help identify and highlight breeders who collect high quality data, help breeders improve their data quality, and increase information available to ram buyers. This research supports Australia's sheep industry to value phenotypes in an objective manner, and can be further expanded to evaluating a flock's data contribution to the reference population.

Acknowledgements

The development of the data quality metrics and DQS has been a long-term project and has received significant funding from MLA and input from Sheep Genetics staff and industry ram breeders. In particular the authors would like to acknowledge the significant contributions of Caris Cornwall (nee Jones), Luke Stephen, Gabrielle Sherring and Andrew Swan.

References

Brown D.J., Tier B and Banks R.B (2001) Proc. Of the 14th AAABG, Queenstown, New Zealand.

Guy S.Z.Y., and Brown D.J. (2021a) Proc. Of the 24th AAABG, Adelaide, Australia.

Guy S.Z.Y., and Brown D.J. (2021b) Proc. Of the 24th AAABG, Adelaide, Australia.

Stephen L.M., Brown D.J., Jones C.E. and Collison, C.E. (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Swan A.A. and Brown D.J. (2007) Proc. Of the 17th AAABG, Armidale, Australia.