Prediction accuracy of direct and indirect approaches, and their relationships with prediction ability of calibration models

T. K. Belay,*[1] B. S. Dagnachew,* S. A. Boison,† and T. Ådnøy*
*Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, PO Box 5003, 1432 Ås, Norway
†Nofima: The Norwegian Institute of Food, Fisheries and Aquaculture Research, Osloveien 1, 1430 Ås, Norway

ABSTRACT

Milk infrared spectra are routinely used for phenotyping traits of interest through links developed between the traits and spectra. Predicted individual traits are then used in genetic analyses for estimated breeding value (EBV) or for phenotypic predictions using a single-trait mixed model; this approach is referred to as indirect prediction (IP). An alternative approach [direct prediction (DP)] is a direct genetic analysis of (a reduced dimension of) the spectra using a multitrait model to predict multivariate EBV of the spectral components and, ultimately, also to predict the univariate EBV or phenotype for the traits of interest. We simulated 3 traits under different genetic (low: 0.10 to high: 0.90) and residual (zero to high: ±0.90) correlation scenarios between the 3 traits and assumed the first trait is a linear combination of the other 2 traits. The aim was to compare the IP and DP approaches for predictions of EBV and phenotypes under the different correlation scenarios. We also evaluated relationships between performances of the 2 approaches and the accuracy of calibration equations. Moreover, the effect of using different regression coefficients estimated from simulated phenotypes (β_p), true breeding values (β_g), and residuals (β_r) on performance of the 2 approaches were evaluated. The simulated data contained 2,100 parents (100 sires and 2,000 cows) and 8,000 offspring (4 offspring per cow). Of the 8,000 observations, 2,000 were randomly selected and used to develop links between the first and the other 2 traits using partial least square (PLS) regression analysis. The different PLS regression coefficients, such as β_p, β_g, and β_r, were used in subsequent predictions following the IP and DP approaches. We used BLUP analyses for the remaining 6,000 observations using the true (co)variance components that had been used for the simulation. Accuracy of prediction (of EBV and phenotype) was calculated as a correlation between predicted and true values from the simulations. The results showed that accuracies of EBV prediction were higher in the DP than in the IP approach. The reverse was true for accuracy of phenotypic prediction when using β_p but not when using β_g and β_r, where accuracy of phenotypic prediction in the DP was slightly higher than in the IP approach. Within the DP approach, accuracies of EBV when using β_g were higher than when using β_p only at the low genetic correlation scenario. However, we found no differences in EBV prediction accuracy between the β_p and β_g in the IP approach. Accuracy of the calibration models increased with an increase in genetic and residual correlations between the traits. Performance of both approaches increased with an increase in accuracy of the calibration models. In conclusion, the DP approach is a good strategy for EBV prediction but not for phenotypic prediction, where the classical PLS regression-based equations or the IP approach provided better results.
Key words: indirect prediction, direct prediction, breeding value, phenotype

INTRODUCTION

Fourier transform mid-infrared (**FT-MIR**) spectrometry is a potential tool for collection of data at population level for phenotypic and genetic analyses of milk components (or other derived traits). An individual's phenotype for a trait is predicted from the FT-MIR spectra. This prediction is dependent on availability of links between the trait of interest and milk spectra. The predicted trait and pedigree information and variance component estimates are used to calculate EBV and other random components included in the model based on a single-trait BLUP approach. Dagnachew et al. (2013b) referred to such an approach as indirect prediction (**IP**) because the multitrait spectral information is not directly used in EBV prediction procedures. Alternatively, genetic analyses can be applied directly on the milk spectral variables or on their factor scores (latent traits). The BLUP predictions of the

random components of the model (EBV, herd test day, permanent environment, and residual) for the traits of interest are then predicted as correlated traits to the corresponding random components of the spectra. Dagnachew et al. (2013b) referred to such an approach as direct prediction (**DP**). Given the strong correlations among milk FT-MIR spectral variables (Soyeurt et al., 2010;Dagnachew et al., 2013a), direct genetic analyses on such correlated spectral variables may result in better accuracy of genetic evaluations (Dagnachew et al., 2013b).

The IP and DP approaches have been used to predict EBV for major milk contents (fat, protein, and lactose) in goats (Dagnachew et al., 2013b) and for traits related to fine milk compositions and technological properties of milk in cows (Bonfatti et al., 2017). Dagnachew et al. (2013b) showed that the DP approach performed better than the IP approach (i.e., relative genetic gain was improved by 3–5% in the DP compared with the IP approach) and also reported high rank correlation coefficients (0.93 to 0.96) between EBV predicted using IP and DP. However, Bonfatti et al. (2017) reported rank correlations ranging from 0.07 to 0.96, but with <0.5 rank correlations for most traits investigated in their study. Belay et al. (2017) adopted the 2 approaches to predict phenotype for BHB in blood from milk spectra and reported a slightly better phenotypic prediction by the IP than the DP approach.

Based on studies done so far, it is difficult to make a conclusive remark on whether the DP approach is better than the IP approach for EBV or phenotype prediction. Each of the studies cited in the preceding paragraph has their limitation. For example, independent chemical analyses (reference values) for the milk contents were not available in the study of Dagnachew et al. (2013b; i.e., the study used phenotypes predicted from the same spectra as reference values for both model calibration and evaluation). Possibly for the reason above, the coefficients of determination (R^2) were very high (>0.96). Moreover, the accuracies of EBV were estimated based on coefficient matrices of the mixed model equations in Dagnachew et al. (2013b). In the study of Bonfatti et al. (2017), reference values measured independently of the spectra were used to develop prediction equations that had medium (0.35) to high (0.86) R^2 values; however, it is difficult to distinguish the approaches that performed better based on that study because the IP and DP approaches were evaluated based on rank correlations. In an attempt to predict phenotypes with IP and DP (Belay et al., 2017), the R^2 were low and data sets used for the model validation as well as for evaluation of the 2 approaches were small.

Furthermore, in the 3 studies, covariance components of the latent traits estimated by the DP approach were converted to variance components to be used in the IP approach using links (regression coefficients) estimated based on phenotypes (β_p). Similarly, EBV of the latent traits were converted into single-trait EBV using phenotype based links. Utilization of a partial least square (**PLS**) regression coefficient estimated from phenotypes (i.e., β_p to convert EBV of latent traits into EBV of trait of interest) does not seem appropriate; this might have an effect on the performance of the approaches. Parameters estimated at one level (e.g., at phenotypic level) were used at another level (e.g., at genetic level); therefore, the effect of using appropriate conversion parameters [e.g., estimated from true breeding values (β_g)] to convert multitrait structures to single-trait structures on performances of the 2 approaches is unknown and needs to be studied. Moreover, relationship between performance of the 2 approaches and accuracy of calibration models is unclear.

Therefore, objectives of our study were (1) to evaluate performance of the IP and DP approaches for prediction of EBV and phenotype under different genetic and residual structures between traits; (2) to evaluate effect of using different PLS regression coefficients (e.g., β_p, β_g, and so on) for converting covariance components or EBV of latent traits into univariate structure on performance of the 2 approaches; and (3) to study the relationship between performance of the 2 approaches and accuracy of calibration models.

MATERIALS AND METHODS

Simulation

A simulation program written in R (R Core Team, 2016) to make single- and multitrait data sets based on pedigree was used. The R codes used for the simulation can be found online (https://github.com/soloboan/Multi-trait_simulations). A base population consisting of 100 sires and 2,000 cows with 3 traits under different genetic and residual correlation scenarios were simulated. Subsequently, 2 generations of data were simulated, with 2,100 parents (100 sires and 2,000 cows) in each generation. It was assumed that a cow would have 4 offspring per generation, resulting in 8,000 offspring per generation, from which parents for the next generation were selected. Sex ratio of offspring was fixed at 50%. Parents were randomly selected and the selected animals were randomly mated by random union of gametes leading to pseudo-overlapping generations as is mostly used in cattle breeding.

Variance components and the corresponding heritabilities used for simulation of the 3 traits are given in Table 1, whereas the different genetic and residual correlation scenarios are presented in Table 2. The