Received: 17 May 2017 | Accepted: 7 October 2017

DOI: 10.1111/jbg.12302

ORIGINAL ARTICLE

WILEY Journal of Animal Breeding and Genetics

A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine litter size traits using Large White and simulated data

A.M. Putz[1] | F. Tiezzi[1] | C. Maltecca[1] | K.A. Gray[2] | M.T. Knauer[1]

[1]Department of Animal Science, North Carolina State University, Raleigh, NC, USA

[2]Smithfield Premium Genetics, Rose Hill, NC, USA

Correspondence
A.M. Putz, Department of Animal Science, North Carolina State University, Raleigh, NC, USA.
Email: putz.austin@gmail.com

Summary

The objective of this study was to compare and determine the optimal validation method when comparing accuracy from single-step GBLUP (ssGBLUP) to traditional pedigree-based BLUP. Field data included six litter size traits. Simulated data included ten replicates designed to mimic the field data in order to determine the method that was closest to the true accuracy. Data were split into training and validation sets. The methods used were as follows: (i) theoretical accuracy derived from the prediction error variance (PEV) of the direct inverse (iLHS), (ii) approximated accuracies from the accf90(GS) program in the BLUPF90 family of programs (Approx), (iii) correlation between predictions and the single-step GEBVs from the full data set ($GEBV_{Full}$), (iv) correlation between predictions and the corrected phenotypes of females from the full data set (Y_c), (v) correlation from method iv divided by the square root of the heritability (Y_{ch}) and (vi) correlation between sire predictions and the average of their daughters' corrected phenotypes (Y_{cs}). Accuracies from iLHS increased from 0.27 to 0.37 (37%) in the Large White. Approximation accuracies were very consistent and close in absolute value (0.41 to 0.43). Both iLHS and Approx were much less variable than the corrected phenotype methods (ranging from 0.04 to 0.27). On average, simulated data showed an increase in accuracy from 0.34 to 0.44 (29%) using ssGBLUP. Both iLHS and Y_{ch} approximated the increase well, 0.30 to 0.46 and 0.36 to 0.45, respectively. $GEBV_{Full}$ performed poorly in both data sets and is not recommended. Results suggest that for within-breed selection, theoretical accuracy using PEV was consistent and accurate. When direct inversion is infeasible to get the PEV, correlating predictions to the corrected phenotypes divided by the square root of heritability is adequate given a large enough validation data set.

KEYWORDS

accuracy comparison, accuracy validation, litter size, piglet mortality, single-step GBLUP

1 | INTRODUCTION

Multiple genomic validation methods for accuracy comparisons have been utilized since the introduction of genomic selection (GS) (Aguilar et al., 2010; Forni, Aguilar, & Misztal, 2011; Saatchi et al., 2011). One of the most common methods is to compare which genomic prediction model performs best in terms of accuracy and bias. Results from these studies have reached a similar conclusion; GS models outperform pedigree-based predictions with little

variability among GS models (Aguilar et al., 2010; Habier, Fernando, Kizilkaya, & Garick, 2011; Tusell, Perez-Rodriguez, Forni, Wu, & Gianola, 2013). In the light of these results, the swine industry has adopted single-step methodology for routine evaluation due to its simplicity and limitations of early marker estimation models (Legarra, Christensen, Aguilar, & Misztal, 2014; Misztal, 2016). Single-step GBLUP (ssGBLUP) was derived to utilize genotyped and non-genotyped individuals in the same BLUP framework by blending the additive relationship matrix (**A**) and genomic relationship matrix (**G**) (Christensen & Lund, 2010; Legarra, Aguilar, & Misztal, 2009).

Much of the focus for GS has been placed on models and far less on the accompanying validation methods being used. Many articles use only one or maybe two methods and have ignored justification of the validation methods being used for accuracy comparisons between models (Abell, Mabry, Dekkers, & Stalder, 2012; Christensen, Madsen, Nielsen, Ostersen, & Su, 2012; Forni et al., 2011; Guo et al., 2015; Lourenco et al., 2014; Serão et al., 2016). In a review paper on GS by de los Campos, Hickey, Pong-Wong, Daetwyler, and Calus (2013), accuracy validation methods were not thoroughly discussed, yet play a critical role to compare alternative breeding programmes. Misztal (2016) discussed validation methods not being thoroughly investigated and stated, "So the quest for the perfect validation continues." There may not be a perfect validation method, but understanding some properties of different methods is warranted. Therefore, the objective of this study was to compare the accuracy of validation methods for swine litter size traits, real and simulated, to determine the optimal method for comparing pedigree BLUP to single-step GBLUP in terms of accuracy and consistency.

2 | MATERIALS AND METHODS

2.1 | Field data

Data provided by Smithfield Premium Genetics (SPG; Rose Hill, NC, USA) were collected from June 2009 through May 2013 for the Large White (LW) purebred maternal breed. It included 8,257 litters from 4,849 LW sows. A complete pedigree was available going back at least three generations. At birth, the number of live born and stillborn piglets was recorded. Litters were recorded and piglets were processed, weighed and cross-fostered within 48 hr of birth. Individual mortality dates were recorded on piglets that died after they were processed. Of the total number of live born piglets, cross-fostering was minimized and occurred for 4.9% of piglets and 17% of litters (see Putz, Tiezzi, Maltecca, Gray, & Knauer, 2015).

Six litter size traits were analysed as follows: total number born (TNB), number born alive (NBA), litter size at day 5 and 10 (LS5, LS10), litter size at weaning (LSW) and number of piglets weaned (NW). All traits were modelled as a trait of the sow. Litter size at day 5, 10 and weaning assigned piglets to their biological litter (Nielsen, Su, Lund, & Madsen, 2013; Putz et al., 2015; Su, Lund, & Sorensen, 2007). In contrast, NW included all piglets that were present in the litter of that sow at the time of weaning, including any piglets cross-fostered onto the litter (i.e., as a nurse dam regardless of biological status).

Animals were genotyped using the Illumina PorcineSNP60 Beadchip. There were a total of 61,565 single nucleotide polymorphisms (SNP) in the raw genotype file. Prior to processing, the number of genotyped animals was 3,264. Software in the BLUPF90 family of programs was utilized for processing (preGSf90). Genotype processing included removing any SNPs with minor allele frequency less than 0.05 or a call rate less than 0.90, mapped to sex chromosomes, and unmapped. Animal genotypes with a call rate less than 0.90 or a parent–progeny conflict were removed. After processing, LW had 32,719 SNPs and 3,195 genotyped animals.

2.2 | Simulated data

Simulated data were created using QMSim (Sargolzaei & Schenkel, 2009) to mimic the field data used in this analysis as closely as possible given a maternal swine-breeding programme. The simulated data were used to help determine the optimal validation method for litter size, given the TBV was available. Ten replicates were analysed to calculate the mean and standard deviation of each method. Historical generations were run for 1,500 generations and simulated 500 sows and 500 boars, followed by an expansion to 1,000 sows and 1,000 boars within ten generations in order to sample enough females for the recent population. These generations were used to generate historical LD between markers and QTL. The recent population included 1,000 females and 50 males per generation and ran for 21 generations. The 21st generation was needed to identify which individuals were selected in generation 20 to obtain a validation phenotype. An important point is that only those selected obtain a phenotype for litter size in swine. Generations 14 through 20 (last seven generations) were selected for data. Only generations 17 through 20 (four generations) were selected as genotyped individuals. Both of these steps were in attempt to get approximately the same number of records and genotyped individuals as seen in the field data. As it was a simulation, all pedigree records were available on all the recent generations, and the pedigree was traced back three generations. The litter size distribution in the simulation was taken from the LSW distribution in the LW population to mimic the same population structure as the real data set. From that, the number of offspring produced for each female would follow that probability mass function. Overlapping generations were

simulated; sows were culled at 80% per generation and boars at 50% per generation based on age. More sows were simulated than were present in the field data, as QMSim does not allow for repeated records. Phenotypes were sex-limited to females.

Another limitation was that QMSim simulates phenotypes for all female offspring when sex-limited. This is an important point that many simulations overlook. In reality, not all females will record a phenotype for litter size, only those selected to be parents in the next generation will generate phenotypes. Therefore, all other phenotypes (those that did not become parents, already sex-limited) were removed to mimic litter size phenotypes in the real data set. For example, if the female population consisted of 100 sows and they each produced exactly 10 offspring, there would be 500 male and 500 female offspring. If culling was 100%, the maximum number of litter size phenotypes obtainable is only 100, not the 500 that would be simulated from QMSim. Random selection was practiced due to this postsimulation data processing step. Selection with EBV from QMSim would also not represent selection accurately because there would be far more phenotypes available than possible with litter size. This could be incorrect if a system had a daughter nucleus or multiplier in which they were feeding back data to the nucleus from these purebred littermates, but it is my understanding this is generally not the case. Even in that case, not all would be selected to move to the next level. Mating between males and females were assigned to minimize inbreeding of the simulated population. The trait simulated had a heritability of 0.10 and a QTL heritability of 0.08, leaving a 0.02 polygenic effect. QTL effects were simulated from a gamma distribution with shape 0.4 (scaled to total QTL variance). Initial allele frequencies were randomly assigned for QTL and markers. Mutation was not simulated for QTL or markers. The phenotypic variance was set to 9.0, approximately the same as litter size (Putz et al., 2015; Su et al., 2007). An important note is that litter size could not be used as the response trait in QMSim, thus litter size was independent of the phenotype being assigned by QMSim. Each sow would randomly be assigned the number of offspring based on the litter size distribution specified (LSW from the LW breed). The response phenotype used was the continuous phenotype assigned by QMSim with approximately the same variance observed for litter size data. It was not the litter size randomly assigned per litter. Litter size was randomly assigned to keep the same population structure in terms of full/half-sib relationships. The resulting simulated data set had a total of 5,700 records (females) and 3,300 genotyped individuals (males and females) after processing. Variance components were estimated within each replicate.

The genome consisted of 18 pairs of chromosomes 100 cM each. Two alleles for 10,000 markers were simulated per chromosome for a total of 180,000 markers. In the last historical generation (1,500), 60,000 markers were randomly selected from those with an allele frequency greater than 0.01. Each chromosome contained 25 QTL with two alleles each for a total of 450 QTL at the beginning of the simulation. On average, 242 QTL remained after the historical populations and 237 were selected for the recent populations. Both markers and QTL were positioned randomly throughout the genome in the first historical generation.

2.3 | Models

Estimated breeding values (EBVs) were calculated using a traditional BLUP prediction model. Analyses were carried out with the BLUPF90 family of programs (Misztal et al., 2002). Models for litter size traits included fixed effects of year-season, farm and parity. Random effects included animal and permanent environmental effects of the dam. The model for simulated data included generation as a fixed effect for the accf90 program and additive genetics of the dam (no repeated measures simulated). The full data set utilizing the pedigree was used to estimate variance components as in Guo et al. (2015). Heritability estimates for litter size traits in the LW ranged from 0.09 to 0.11 (Putz et al., 2015). The pedigree was traced back three generations.

Single-step GBLUP was implemented in this analysis (Christensen & Lund, 2010; Legarra et al., 2009). The computation for $\mathbf{H^{-1}}$ (Lourenco et al., 2014) was

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A^{-1}}$ is the inverse of the numerator relationship matrix (**A**) including all animals; **G** is the genomic relationship matrix (from default preGSf90, VanRaden, 2008); $\mathbf{A}_{22}^{-1}$ is the inverse of the **A** matrix for only genotyped animals. The parameter values were $\alpha = 0.95$, $\beta = 0.05$, $\tau = 1$ and $\omega = 1$. Christensen et al. (2012) showed very small differences between the standard **H** and adjusted **H**, so there was no attempt to adjust the matrix. Observed allele frequencies were used to centre and scale the observed genotype matrix.

2.4 | Validation methods

Forward validation was used in this analysis. In the field data, the last 3 year-seasons were masked as the validation data set, and predictions were made using the first 13 year-seasons as the training data set. The validation data set contained 29% of the observations for Large White. The last 3 year-seasons were selected because the goal in a breeding programme should be to predict breeding values

of young replacements in the next generation, not multiple generations. In the simulated data, generation 20 was masked, and predictions were made from generations 14 through 19 (six generations) to approach a similar data set size as the field data.

Four general methods were used to calculate accuracy: using prediction error variance (PEV) from the inverse left-hand side (iLHS) and calculating the traditional accuracy (Forni et al., 2011), approximated accuracies (Misztal & Wiggans, 1988; Misztal et al., 2013), correlating predictions to the GEBV from the full data set (similar to Aguilar et al., 2010) and correlating predictions to the corrected phenotypes (Abell et al., 2012; Christensen et al., 2012; Lourenco et al., 2014). There were three separate implementations using the corrected phenotypes (see below). For simulated data, the true breeding values (TBV) were known and correlated to the genomic and pedigree predictions to obtain the true accuracy of both models. All of the accuracies calculated were only within the validation data set.

First, the iLHS (direct inverse) was calculated to obtain the standard error of prediction (SEP) to calculate the traditional accuracy. Fortran FSPAK (package for matrix inversions) solving method was used in the BLUPF90 program (i.e., these were not approximations). The equation from Mrode (2014) that was used to calculate traditional accuracy was

$$r_i = \sqrt{1 - \left(\mathrm{SEP}_i^2/\sigma_a^2\right)},$$

where i = 1, 2, …, number of animals in validation; SEP is the SEP output from BLUPF90; and σ_a^2 is the additive genetic variance of the trait being analysed. Accuracy was defined as the mean of the vector for genotyped sows in the validation set.

Second, approximations from the accf90 or accf90(GS) program in the BLUPF90 family of packages were calculated (Approx, Misztal & Wiggans, 1988; Misztal et al., 2013). It is unknown what method was implemented from Misztal et al. (2013), as it is proprietary software (it has been updated more recently). Year-season was used as the major fixed effect in the approximations for the field data set. In simulated data, generation was used as the only fixed effect for approximations. As for iLHS, accuracy was defined as the mean of the vector for genotyped sows in the validation set.

Third, single-step GBLUP predictions from the full data set ($\mathrm{GEBV}_{\mathrm{Full}}$, i.e., not split between training and validation) were used to correlate predictions (Aguilar et al., 2010). Accuracy was defined as

$$r = \mathrm{cor}\left((\mathrm{G})\mathrm{EBV}_t,\ \mathrm{GEBV}_{\mathrm{Full}}\right),$$

where $(\mathrm{G})\mathrm{EBV}_t$ was either the GEBV from the ssGBLUP prediction ($\mathbf{H}^{-1}$) or the EBV from the BLUP pedigree ($\mathbf{A}^{-1}$)-based prediction from the training data set for validation animals. The correlation was carried out only for the validation animals in both the field data set and the simulated data set. This was in an attempt to mimic a method used by Aguilar et al. (2010) where he used EBV_{09} (full data set, corresponding to year 2009) as the response in a model and predictions from EBV_{04} (training data).

Fourth, the corrected phenotypes (Y_c) from the full data set adjusted for fixed effects (parity, farm and year-season) were used as the response to be predicted. Accuracy was defined as in Lourenco et al. (2014) as

$$r = \mathrm{cor}\left((\mathrm{G})\mathrm{EBV}_t,\ Y_c\right),$$

where Y_c was the corrected phenotypes for sows from the full data set; $(\mathrm{G})\mathrm{EBV}_t$ as previously defined. Three different implementations using corrected phenotypes were used. They are: (i) the correlation between the (G)EBV and the corrected phenotypes of the genotyped sows in validation with no records in the training data set (Y_c; Lourenco et al., 2014), (ii) the correlation from (i) divided by the square root of heritability (Y_{ch}; similar to Guo et al., 2015; Serão et al., 2016) and (iii) the correlation in the validation data set between the breeding value prediction of genotyped sires and the average of their daughters' corrected phenotypes (Y_{cs}; Abell et al., 2012). Sires had to have at least five daughter phenotypes in validation. Daughters used to calculate the average corrected phenotype had no phenotypes in the training data set, only from the validation data set.

Any sow with a record in the training data set was removed from the validation data set (left in the training data). This resulted in 194 genotyped LW sows. There were 34 genotyped sires with at least five daughters in the validation data set. For the simulated data, 800 genotyped females and 25 genotyped males were used for validation.

3 | RESULTS

3.1 | Field data

Theoretical accuracies from iLHS were consistent across traits (*SD* ranging from 0.01 to 0.03) and showed an improvement from 0.27 to 0.37 for EBV and GEBV selection in LW (Table 1). This was an increase of 37%. Approximated accuracies showed only a slight improvement from pedigree predictions (0.41 to 0.43) and were slightly higher than iLHS with very little variation (~0.01). Correlating predictions to $\mathrm{GEBV}_{\mathrm{Full}}$ yielded very high accuracies for ssGBLUP (0.35 to 0.79), but were consistent across traits. Corrected phenotypes methods tended to show more variation than the other methods and underestimated the iLHS accuracies (0.04 to 0.27). The most amount of variation was observed for Y_{cs}.

TABLE 1 Accuracies (*SD*) averaged across trait for single-step GBLUP (GEBV) and pedigree (EBV) predictions from the training data set in Large White population (n = 194, n = 34 for Y_{cs})

	EBV	GEBV
iLHS[a]	0.27 (0.01)	0.37 (0.03)
Approx[b]	0.41 (0.01)	0.43 (0.00)
$GEBV_{Full}$[c]	0.35 (0.05)	0.79 (0.04)
Y_c[d]	0.04 (0.03)	0.09 (0.04)
Y_{ch}[e]	0.12 (0.08)	0.27 (0.12)
Y_{cs}[f]	0.09 (0.17)	0.18 (0.15)

[a]Average theoretical accuracy calculated from the PEV from the direct inverse of LHS.
[b]Average accuracy from approximations obtained from the accf90(GS) (genomic selection) program.
[c]Correlation between single-step GEBV predictions from the full data set and predictions from the training data set.
[d]Correlation between corrected phenotypes of dams and predictions.
[e]Y_c divided by the square root of the heritability.
[f]Correlation between average corrected phenotypes of daughters for sires with at least five daughters in validation and predictions.

Each validation method's accuracy across trait was presented in a separate facet (validation method) with the y-scale fixed across all methods within figure (Figure 1). Estimates for iLHS, Approx and $GEBV_{Full}$ were higher for GEBV than EBV for all six traits. For these three methods, the increase ranged from 0.08 to 0.13, 0.02 to 0.02 and 0.41 to 0.48, respectively. These three methods also showed less variation across traits, although some variation was observed. Approximations were only slightly higher than the iLHS accuracies with GEBV slightly outperforming EBV. Three corrected phenotype methods (Y_c, Y_{ch} and Y_{cs}) were more variable and had one trait reranked each between GEBV and EBV. NW was reranked for Y_c and Y_{ch}, while TNB was reranked for Y_{cs}. The range of increase for Y_c, Y_{ch} and Y_{cs} was −0.02 to 0.09, −0.04 to 0.32 and −0.03 to 0.15, respectively.

3.2 | Simulated data

Correlations between TBV with EBV and GEBV (i.e., true accuracies) were 0.34 and 0.44, respectively (Table 2). The increase was the same (0.10) as the field data set. This was a 29% increase for ssGBLUP on average. Averaged across replicate, the Y_{ch} method performed the best in terms of being the closest to the TBV correlations (difference of 0.02 and 0.01 for EBV and GEBV, respectively). As expected, the iLHS was also very close to the true accuracy with a difference of 0.04 for EBV predictions and 0.02 for GEBV predictions. As in the field data, the $GEBV_{Full}$ method showed high accuracies of 0.56 for EBV and 0.82 for GEBV. Variability across replicate was much higher for

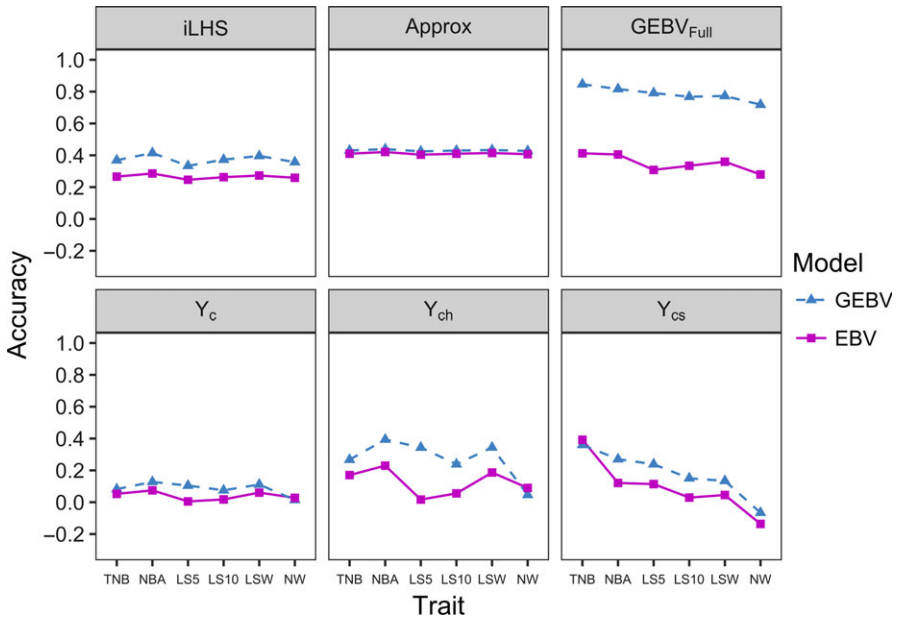

FIGURE 1 Accuracies across trait for single-step GBLUP (GEBV) and pedigree (EBV) predictions from the training data set in the Large White population (n = 194, n = 34 for Y_{cs}). TNB—total number born, NBA—number born alive, LS5—litter size at day 5, LS10—litter size at day 10, LSW—litter size at weaning, NW—number weaned. iLHS—average theoretical accuracy calculated from the PEV from the direct inverse of LHS, Approx—average accuracy from approximations obtained from the accf90(GS) program, $GEBV_{Full}$—correlation between single-step GEBV predictions from the full data set and predictions from the training data set, Y_c—correlation between corrected phenotypes of dams and predictions, Y_{ch}—Y_c divided by the square root of the heritability, Y_{cs}—correlation between average corrected phenotypes of daughters for sires with at least five daughters in validation and predictions. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Accuracies (*SD*) averaged across replicate from simulated data using single-step GBLUP (GEBV) and pedigree (EBV) predictions in the training data set (n = 800, n = 25 for Y_{cs})

	EBV	GEBV
TBV[a]	0.34 (0.06)	0.44 (0.06)
iLHS[b]	0.30 (0.02)	0.46 (0.01)
Approx[c]	0.34 (0.00)	0.37 (0.00)
$GEBV_{Full}$[d]	0.56 (0.06)	0.82 (0.03)
Y_c[e]	0.11 (0.05)	0.14 (0.03)
Y_{ch}[f]	0.36 (0.15)	0.45 (0.11)
Y_{cs}[g]	0.14 (0.22)	0.17 (0.22)

[a]Correlation of predictions to true breeding values.
[b]Average theoretical accuracy calculated from the PEV from the direct inverse of LHS.
[c]Average accuracy from approximations obtained from the accf90(GS) program.
[d]Correlation between single-step GEBV predictions from the full data set and predictions from the training data set.
[e]Correlation between corrected phenotypes of dams and predictions.
[f]Y_c divided by the square root of the heritability.
[g]Correlation between average corrected phenotypes of daughters for sires with at least five daughters in validation and predictions.

the Y_{cs} method than any other, followed by Y_{ch} and Y_c. Both Y_{cs} and Y_c were less than half of the true accuracies (0.11 to 0.17). Approximations were consistent with the field data, showing almost no variation across replicates and a slight increase in accuracy for GEBV predictions.

Correlations between predictions and TBV were higher for GEBV than EBV for all but replicate six (Figure 2). The increase in accuracy ranged from 0 to 0.16 for TBV. Approximations were the most consistent followed by iLHS, $GEBV_{Full}$, Y_c/Y_{ch} and Y_{cs}. Corrected phenotype methods performed better than in the field data, most likely due to a larger validation size (n = 800). In eight of the ten replicates, GEBV outperformed the EBV predictions for Y_c and Y_{ch}. It is worth noting that Y_c and Y_{ch} closely followed the true accuracy correlations by replicate. For instance, replicates 3 and 6 accuracy values showed little increase for GEBV in both TBV and Y_{ch}, while in replicates 7–10, the increase was higher for both methods. The Y_{cs} method showed a dramatic amount of variability and did not seem to mimic the true accuracies well. In six of the ten replicates, the EBV predictions outperformed the GEBV predictions. This method included only 25 sires, which most likely caused much of the variation observed.

4 | DISCUSSION

Many methods have been used in genomic accuracy validation; this research begins to address some of the pros and cons of using different accuracy validation methods. Results from the simulation were in general consistent with accuracies calculated in the field data set for litter size traits. These simulated data validated the effectiveness of using the iLHS and Y_{ch} to estimate the true accuracy, which were more consistent across trait and replicate. However, this was done for within-breed selection in a purebred line. Be careful not to extrapolate results to other scenarios without further validation (e.g., validate across lines/breeds or crossbred performance). This should be a direction for future research. Also be careful not to extrapolate to other species or traits without further investigation. Results may be sensitive to population structure (number of full/half-sib families) or the trait in terms of when data are collected (i.e., growth and feed intake prior to selection and litter size after selection). The corrected phenotype methods showed more variability in terms of reranking of models across traits and replicates so care should be taken interpreting these results with small sample sizes.

Accuracies were quite high for the $GEBV_{Full}$ method and turned out to be a very poor method for accuracy validation for swine litter size traits. This was expected because of extensive overlapping information (not independent) between the GEBV predictions from the training data set and the predictions from the full data set. Only one generation of data was added to these predictions for a lowly heritable trait. This was in contrast to Aguilar et al. (2010) in which 5 years of data were covered by the validation data set. This could span multiple generations for cows and could decrease the accuracy estimated. Accuracies reported in terms of reliability were 0.50 for the best model, which corresponds to an accuracy of 0.71 (Aguilar et al., 2010). This method may be better to measure the stability of the models used and/or better utilized in dairy cattle because dairy bulls can have a very high accuracy with a large number of progeny in the validation population. It is not recommended for litter size accuracy in swine based on the data presented (see tables and figures).

Results showed the Y_c method underestimated the true accuracy in simulated data and the iLHS accuracy in the real data set. The Y_c method may help with model comparisons (**H** versus **A**), but lacks the ability to estimate the true accuracy in terms of absolute value without dividing by the square root of heritability. The Y_{ch} method was successful at estimating the true accuracy in simulated data. Abell, Dekkers, Rothschild, Mabry, and Stalder (2014) showed that adoption of GS would depend on the *true* increase in accuracy and economic considerations of logistics, data collection/analysis and cost of genotyping. For example, in the LW, the increase in accuracy was 125% (from 0.04 to 0.09) using the Y_c method. In contrast, the increase in accuracy for iLHS method was only 37% in LW (from 0.27 to 0.37). Furthermore, the absolute value is too small from the Y_c method to make a significant impact on the decision to adopt GS as observed in the LW

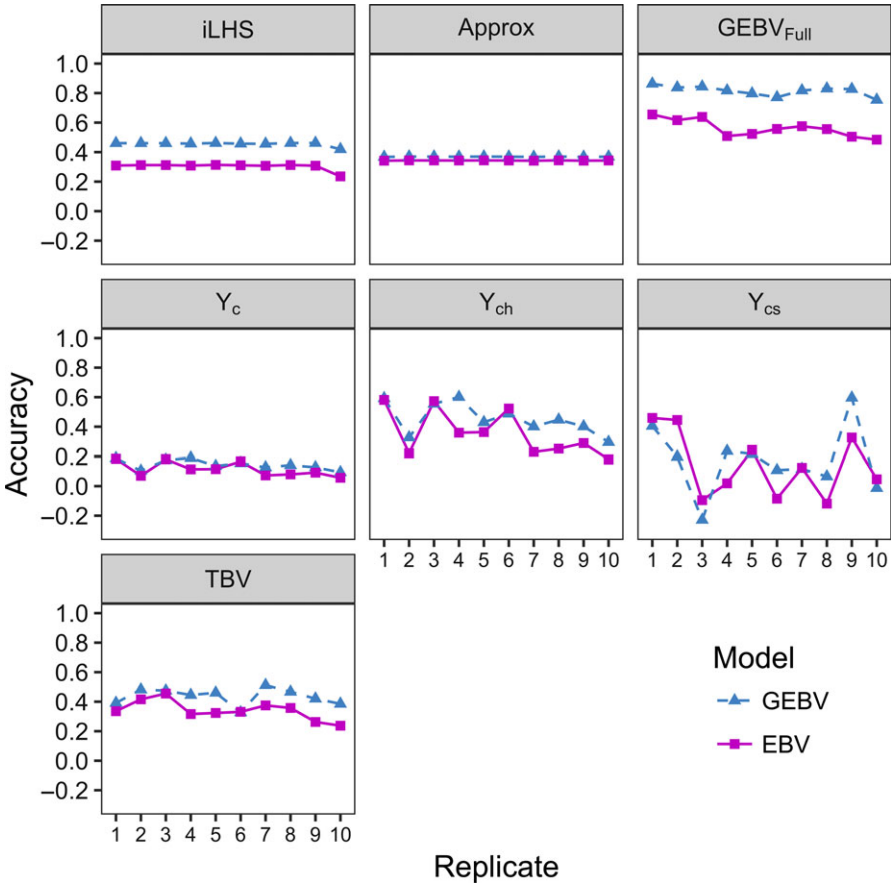

FIGURE 2 Accuracies across replicate for the simulated data sets comparing single-step GBLUP (GEBV) and pedigree (EBV) predictions from the training data set (n = 800, n = 25 for Y_{cs}). iLHS—average theoretical accuracy calculated from the PEV from the direct inverse of LHS, Approx—average accuracy from approximations obtained from the accf90 (GS) program, $GEBV_{Full}$—correlation between single-step GEBV predictions from the full data set and predictions from the training data set, Y_c—correlation between corrected phenotypes of dams and predictions, Y_{ch}—Y_c divided by the square root of the heritability, Y_{cs}—correlation between average corrected phenotypes of daughters for sires with at least five daughters in validation and predictions, TBV—correlation of predictions to true breeding values. [Colour figure can be viewed at wileyonlinelibrary.com]

population in the current study. Low values such as 0.09 for the Y_c accuracy in the LW make it appear insignificant and might result in other programmes waiting to adopt GS while genotyping costs come down. This should also be considered as the swine industry considers replacing the current SNP panel with higher density SNP panels or whole-genome sequencing (see Hickey, 2013) as well as continual comparison of GS models.

Using corrected phenotypes of sows was expected to perform poorly when the heritability is low. The R^2 for litter size models was ~0.05 or below. For this reason, daughter residuals were averaged for sires (Y_{cs}) in an attempt to negate the extra noise and regain more of the genetic merit. This was a simple solution to evaluate sires in the past (Thompson, 1979). In the preliminary analysis, it was discovered that a minimum number of daughter records were needed to obtain better accuracies. Abell et al. (2012) used a cut-off of ten daughters, but given the limited sample size of sires in the current study for validation, five was chosen. This balance between increasing the number of daughters averaged and sample size needs to be considered. In the LW, there were 160 more genotyped sows than genotyped sires with at least five daughters in validation. As with Y_c or Y_{ch}, results suggest using Y_{cs} with a limited number in validation will result in highly variable results and should be considered prior to analysis. In general, pure-bred lines in swine will suffer from a small number of sires used in a validation data set leading to the method performing poorly.

It was observed in the LW population that iLHS, approximations and $GEBV_{Full}$ methods all showed that GEBV predictions outperformed the EBV predictions. In contrast, Y_c showed that EBV and GEBV performed equally well for NW in LW, for example. This may weigh in the decision to adopt one litter size trait over another depending on which validation method is chosen.

There still does not seem to be a standard among researchers for selecting a **G** matrix and blending the **G** and **A** matrices. However, the default for the BLUPF90 programs is to use observed allele frequencies for centring and scaling and a weight of 0.95 and 0.05 for **G** and **A**, respectively (i.e., α = 0.95, β = 0.05). Results from Forni et al. (2011) show that GEBV estimates from different **G** matrices resulted in extremely high correlations among one another (~0.99 or above). Therefore, no attempt was made to utilize different relationship matrices in the current study, as the accuracy wouldn't change for methods such as Y_c. Forni et al. (2011) reported an accuracy increase from 0.22 to 0.30 using the observed frequencies for **G** using the PEV. Other methods for obtaining **G** have been compared, and the results for theoretical accuracy seemed

inflated for several of them (Forni et al., 2011). Lourenco et al. (2014) used 0.7, 0.3, 0.7 and 0.8 for α, β, τ and ω, corresponding to the weights for **G**, $\mathbf{A_{22}}$, $\mathbf{G^{*-1}}$ and $\mathbf{A_{22}^{-1}}$, respectively. These weights are not expected to change the accuracy much and could be used to increase stability, but it could change based on what validation method is used.

Forward validation was used as opposed to a common cross-validation technique. Cross-validation can lead to very high, unrealistic estimates of accuracies (Misztal, 2016). The goal of GS is to have the ability to genotype at a young age prior to phenotyping and predict breeding values accurately. Cross-validation can be useful to predict true accuracy from predictive ability (Legarra, Manfredi, & Elsen, 2008), but does not mimic a real breeding programme. This study used highly correlated traits (litter size, see Putz et al., 2015) to obtain a degree of replication and replicates in simulation.

It is highly desirable that the method used to validate be stable and predict the true accuracy well in terms of absolute value. The biggest unanswered question will be how these validation method results change with the trait (genetic architecture) and species (population structure). Data from the simulation had to be processed to mimic the true phenotypes in a swine-breeding programme. Only animals that became parents for the next generation can generate phenotypes for litter size. Traits such as average daily gain should have phenotypes for most offspring whether they become parents or not, thus changing the advantage for GS, unless preselection is needed for traits such as feed intake. Carcass traits will also have a different level of phenotyping depending on the breeding programme. This may change the results from the current study because relatively few full and half-sibs would also have phenotypes for litter size and are expected to have lower accuracy. More research could also be completed, extending to estimating bias, as the true accuracies are never known to regress on. No method such as iLHS exists for bias, which increases the need to find an alternative method even if the inversion of LHS is feasible for accuracy.

One important discovery of this research was that the data set size in validation had an impact on accuracies, especially with the corrected phenotype methods. In an initial analysis, Landrace was included but had a very limited number of genotyped individuals in validation so accuracies from Y_c, Y_{ch} and Y_{cs} fluctuated dramatically and were subsequently removed due to the results (many were over 1, the upper bound for accuracy). LW had 194 genotyped sows for validation and still seemed to show a large amount of variation compared to the simulated data. However, iLHS and Approx were not affected greatly in Landrace compared to Large White, leading to about the same increase in accuracy. Future research may be needed to address how large the validation population needs to be so these correlations are stable.

5 | CONCLUSIONS

When inverting LHS is possible, it appears to be the best validation method available when using ssGBLUP, in terms of stability and accuracy within population. Results of this study suggest that the theoretical accuracy from inverse LHS was very consistent and approximated the true accuracy in the simulated data well. With simulated data, the correlation between corrected phenotypes of dams and the predictions divided by the square root of heritability performed very well averaged across replicate. With field data it did not perform as well when compared to iLHS. It is important to ensure that enough genotyped validation animals exist if using corrected phenotypes compared to other methods due to instability of this method. Breeding programmes need to be careful which validation method they choose and should investigate multiple methods if possible. Future research is needed to address this topic across different traits (architectures) and species, as they are expected to change based on the trait and/or population.

ORCID

A.M. Putz http://orcid.org/0000-0001-6459-7123

REFERENCES

Abell, C. E., Dekkers, J. C. M., Rothschild, M. F., Mabry, J. W., & Stalder, K. J. (2014). Total cost of estimation for implementing genome-enabled selection in a multi-level swine production system. *Genetics Selection Evolution*, *46*, 32. https://doi.org/10.1186/1297-9686-46-32

Abell, C. E., Mabry, J. W., Dekkers, J. C. M., & Stalder, K. J. (2012). Relationship between litters per sow per year sire breeding values and sire progeny means for farrowing rate, removal parity and lifetime born alive. *Journal of Animal Breeding and Genetics*, *130*, 64–71. https://doi.org/10.1111/j.1439-0388.2012.01012.x

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*, 743–752. https://doi.org/10.3168/jds.2009-2730

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*, 327–345. https://doi.org/10.1534/genetics.112.143313

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetic Selection Evolution*, *42*, 2. https://doi.org/10.1186/1297-9686-42-2

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T., & Su, G. (2012). Single-step methods for genomic evaluation in pigs. *Animal*, *6*(10), 1565–1571. https://doi.org/10.1017/S1751731112000742

Forni, S., Aguilar, I., & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree, and genomic information. *Genetics Selection Evolution*, *43*, 1. https://doi.org/10.1186/1297-9686-43-1

Guo, X., Christensen, O. F., Ostersen, T., Wang, Y., Lund, M. S., & Su, G. (2015). Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. *Journal of Animal Science*, *93*, 503–512. https://doi.org/10.2527/jas.2014-8331

Habier, D., Fernando, R. L., Kizilkaya, K., & Garick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*, 186. https://doi.org/10.1186/1471-2105-12-186

Hickey, J. M. (2013). Sequencing millions of animals for genomic selection 2.0. *Journal of Animal Breeding and Genetics*, *130*, 331–332. https://doi.org/10.1111/jbg.2013.130.issue-5

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*, 4656–4663. https://doi.org/10.3168/jds.2009-2061

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single step, a general approach for genomic selection. *Livestock Science*, *166*, 54–65. https://doi.org/10.1016/j.livsci.2014.04.029

Legarra, A. C. Robert-Granie, Legarra, A. , Robert-Granie, C. , & Elsen, J. (2008). Performance of genomic selection in mice. *Genetics*, *180*, 611–618. https://doi.org/10.1534/genetics.108.088575

Lourenco, D. A. L., Misztal, I., Tsuruta, S., Aguilar, I., Lawlor, T. J., Forni, S., & Weller, J. I. (2014). Are evaluations on young genotyped animals benefiting from the past generations. *Journal of Dairy Science*, *97*, 3930–3942. https://doi.org/10.3168/jds.2013-7769

Misztal, I. (2016). Is genomic selection now a mature technology? *Journal of Animal Breeding and Genetics*, *133*, 81–82. https://doi.org/10.1111/jbg.2016.133.issue-2

Misztal, I., Tsuruta, S., Aguilar, I., Legarra, A., VanRaden, P. M., & Lawlor, T. J. (2013). Methods to approximate reliabilities in single-step genomic evaluation. *Journal of Dairy Science*, *96*, 647–654. https://doi.org/10.3168/jds.2012-5656

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., & Lee, D. H. (2002). BLUPF90 and related programs (BGF90). Communication No. 28-07 in Proc. 7th World Cong. Genet. Appl. Livest. Prod., Montpellier, France.

Misztal, I., & Wiggans, G. R. (1988). Approximation of prediction error variance in large-scale animal models. *Journal of Dairy Science*, *71*(Suppl 2), 27–32. https://doi.org/10.1016/S0022-0302(88)79976-2

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values*, 3rd ed. Oxfordshire, UK: CABI. https://doi.org/10.1079/9781780643915.0000

Nielsen, B., Su, G., Lund, M. S., & Madsen, P. (2013). Selection for increased number of piglets at d 5 after farrowing has increased litter size and reduced piglet mortality. *Journal Animal Science*, *91*, 2575–2582. https://doi.org/10.2527/jas.2012-5990

Putz, A. M., Tiezzi, F., Maltecca, C., Gray, K. A., & Knauer, M. T. (2015). Variance component estimates for alternative litter size traits in swine. *Journal of Animal Science*, *93*, 5153–5163. https://doi.org/10.2527/jas.2015-9416

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, *25*, 680–681. First published January 28, 2009. https://doi.org/10.1093/bioinformatics/btp045

Saatchi M., McClure M. C., McKay S. D., Rolf M. M., Kim J., Decker J. E., Taxis T. M., Chapple R. H., Ramey H. R., Northcutt S. L., Bauck S., Woodward B., Dekkers J. C. M., Fernando R. L., Schnabel R. D., Garrick D. J., & Taylor J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, *43*, 40. https://doi.org/10.1186/1297-9686-43-40

Serão, N. V. L., Kemp, R. A., Mote, B. E., Willson, P., Harding, J. C. S., Bishop, S. C., … Dekkers, J. C. M. (2016). Genetic and genomic basis of antibody response to porcine reproductive and respiratory syndrome (PRRS) in gilts and sows. *Genetics Selection Evolution*, *48*, 51. https://doi.org/10.1186/s12711-016-0230-0

Su, G., Lund, M. S., & Sorensen, D. (2007). Selection for litter size at day five to improve litter size at weaning and piglet survival rate. *Journal of Animal Science*, *85*, 1385–1392. https://doi.org/10.2527/jas.2006-631

Thompson, R. (1979). Sire evaluation. *Biometrics*, *35*, 339–353. https://doi.org/10.2307/2529955

Tusell, L., Perez-Rodriguez, P., Forni, S., Wu, X. L., & Gianola, D. (2013). Genome-enabled methods for predicting litter size in pigs: A comparison. *Animal*, *7*(11), 1739–1749. https://doi.org/10.1017/S1751731113001389

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*, 4414–4423. https://doi.org/10.3168/jds.2007-0980

How to cite this article: Putz AM, Tiezzi F, Maltecca C, Gray KA, Knauer MT. A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine litter size traits using Large White and simulated data. *J Anim Breed Genet*. 2018;135:5–13. https://doi.org/10.1111/jbg.12302