Ni, Z., Dekkers, J., and Fernando, R. (2026) A statistical test to compare the accuracy of genetic predictions, Proceedings 13th World Congress on Genetics Applied to Livestock Production 24312: https://doi.org/10.31274/wcgalp.24312 © 2026 *Ni, et al.* CC BY-NC-ND

A statistical test to compare the accuracy of genetic predictions

Z. Ni[1*], J. Dekkers[1], and R. Fernando[1]

[1]*Iowa State University, Ames, IA, USA;* [*]*Corresponding author: zqni@iastate.edu*

Introduction

Correlation-based validation is a standard way to report genomic prediction performance. In breeding, the accuracy of a predictor $\mathbf{x}_k$ is its Pearson correlation with the unobserved true breeding value $\mathbf{u}$, i.e., $\mathrm{corr}(\mathbf{u}, \mathbf{x}_k)$,. In validation, a common proxy is the predictive ability (PA), $\mathrm{corr}(\mathbf{y}, \mathbf{x}_k)$, where $\mathbf{y}$ and $\mathbf{x}_k$ are vectors of phenotypes and the k^{th} set of estimated breeding values (EBV), which is usually divided by the square root of trait heritability to derive EBV accuracy. Therefore, testing equality of accuracies of two sets of EBV for a trait, $\mathbf{x}_1$ and $\mathbf{x}_2$, reduces to testing the null hypothesis $\mathrm{corr}(\mathbf{y}, \mathbf{x_1}) = \mathrm{corr}(\mathbf{y}, \mathbf{x_2})$. The Williams t-test (Williams, 1959; Steiger, 1980) is widely used for testing the equality of correlations of two variables with a common response $\mathbf{y}$. Recent analytical frameworks for comparing coefficients of determination, such as r2redux (Momin et al., 2023), have advanced the field by providing rigorous confidence intervals. However, all these tests assume that elements of $\mathbf{y}$ are independent and identically distributed (i.i.d). This assumption is severely violated in breeding populations, which can render tests to have inflated Type I errors (T1E). Dependence also exists among predictors $\mathbf{x}_k$, arising from shared training data and model structures but this is model and data specific and often intractable. Consequently, we treat the validation predictors as realized fixed covariates and focus only on accounting for the structured dependence in $\mathbf{y}$. This motivated our development and evaluation of a generalized least squares wild bootstrap (GLS-WB) as a dependence-aware paired test for comparing validation correlations in related cohorts via a GRM (genomic relationship matrix)-based working covariance model for $\mathbf{y}$ to perform inference conditional on the realized predictors in validation.

Materials & Methods

The proposed GLS-WB results in a two-sided p-value for the difference in PA of two realized predictors for the same validation set $\mathbf{y}_{\mathrm{val}}$. The key steps are: (i) use a working covariance matrix $\mathbf{V}$ for the validation phenotypes to decorrelate $\mathbf{y}_{\mathrm{val}}$; (ii) fit the model restricted to the null hypothesis of equal correlations (H_0) to obtain residuals under H_0 ($\hat{\mathbf{e}}_w$), with dependence reduced based on the working model; and (iii) apply a restricted-residual wild bootstrap by randomly flipping the sign (+/-) of each decorrelated $\hat{\mathbf{e}}_w$ (Davidson and Flachaire, 2008) to generate $\mathbf{y}_{\mathrm{val}}$ under H_0 of no difference of its correlation with the two sets of EBVs, resulting in an empirical distribution for the difference in PA under H_0.

Hypothesis and test statistic. Let $\mathbf{y}_{\mathrm{val}}$ denote the vector of validation phenotypes and let $\mathbf{x}_1, \mathbf{x}_2$ denote the two realized predictor vectors (EBVs) for the validation individuals. We define PA of the k^{th} predictor as $r_k = \mathrm{corr}(\mathbf{y}_{\mathrm{val}}, \mathbf{x}_k)$ and the gap as $\Delta r = r_1 - r_2$. Given $\mathbf{x}_1$ and $\mathbf{x}_2$, we testH_0: $\Delta r = 0$. To approximate the covariance of the validation phenotypes based on relatedness, we use an additive genetic model with working covariance $\mathrm{cov}(\mathbf{y}_{\mathrm{val}}) \approx \mathbf{V} = h^2\mathbf{G}_{\mathrm{val}} + (1 - h^2)\mathbf{I}$, where $\mathbf{G}_{\mathrm{val}}$ is the GRM among validation individuals and h^2 is heritability. With $\mathbf{L}\mathbf{L}^{\mathrm{T}}$ as the Cholesky factorization of $\mathbf{V}$, elements of $\mathbf{y}_{\mathrm{val}}$ can be decorrelated as $\mathbf{y}_w = \mathbf{L}^{-1}\mathbf{y}_{\mathrm{val}}$, which is also applied to any validation-set vector when needed (e.g., $\mathbf{s}_w = \mathbf{L}^{-1}\mathbf{s}$).
After standardizing predictors by their mean and standard deviation, $\breve{\mathbf{x}}_k$, shared and contrast direction vectors are defined as $\mathbf{s}=(\breve{\mathbf{x}}_1+\breve{\mathbf{x}}_2)/\sqrt{2}$ and $\mathbf{d}=(\breve{\mathbf{x}}_1-\breve{\mathbf{x}}_2)/\sqrt{2}$. In decorrelated space, we consider the working regression model $\mathbf{y}_w = \beta_0\mathbf{1}_w + \beta_S\mathbf{s}_w + \beta_D\mathbf{d}_w + \boldsymbol{\varepsilon}_w$, where $\mathbf{y}_w$ denotes

the decorrelated validation phenotype; $\mathbf{1}_w$, $\mathbf{s}_w$, and $\mathbf{d}_w$ are decorrelated versions of the intercept and of the shared, and contrast direction vectors; β_0, β_S, and β_D are unknown regression coefficients; and $\boldsymbol{\varepsilon}_w$ is the vector of iid error terms. For this model, $\Delta r=0$ corresponds to no association of the contrast direction vector $\mathbf{d}_w$ with $\mathbf{y}_w$, i.e., H_0: $\beta_D=0$. Thus, the null model uses only the intercept and the shared direction vector $\mathbf{s}_w$, with incidence matrix $\mathbf{W}_0=[\mathbf{1}_w, \mathbf{s}_w]$. Considering the QR decomposition of $\mathbf{W}_0=\mathbf{Q}_0\mathbf{R}_0$, with $\mathbf{Q}_0^\top\mathbf{Q}_0=\mathbf{I}$, fitting the restricted model in decorrelated space results in fitted values $\hat{\mathbf{y}}_{w,0}=\mathbf{Q}_0\mathbf{Q}_0^\top\mathbf{y}_w$ and residuals $\hat{\mathbf{e}}_w=\mathbf{y}_w-\hat{\mathbf{y}}_{w,0}$, which are used as the basis for a bootstrap.

For each bootstrap draw b, we draw a vector of random sign multipliers $\mathbf{v}^{(b)} \in \{-1,+1\}^{n_{val}}$ for all n_{val} validation individuals and generate $\mathbf{y}_w^{(b)} = \hat{\mathbf{y}}_{w,0} + (\mathbf{I}-\mathbf{Q}_0\mathbf{Q}_0^\top)(\hat{\mathbf{e}}_w \odot \mathbf{v}^{(b)})$. Here, the projection $(\mathbf{I}-\mathbf{Q}_0\mathbf{Q}_0^\top)$ ensures that $\hat{\mathbf{e}}_w \odot \mathbf{v}^{(b)}$ remains in the orthogonal complement of the null subspace. After mapping back to the original scale via $\mathbf{y}_{val}^{*(b)} = \mathbf{L}\mathbf{y}_w^{*(b)}$, a sample of Δr under H_0 is computed as $\text{corr}\left(\mathbf{y}_{val}^{*(b)}, \mathbf{x}_1\right) - \text{corr}\left(\mathbf{y}_{val}^{*(b)}, \mathbf{x}_2\right)$. A two-sided p-value for the test is computed as the proportion of bootstrap replicates for which $|\Delta r| \geq$ the observed $|\Delta r|$.

Simulation design to evaluate GLS-WB. We used genotype and pedigree data from Cleveland *et al.* (2012), with 3,534 pigs genotyped for 52,843 SNPs after quality control. To mimic a forward prediction scenario, we split the data into a training set of older animals and a validation set of younger animals based on pedigree depth. For each set, we randomly sampled half to obtain a training dataset of 989 records and animals and a validation dataset of 778 animals to reach relevant realized accuracies and reduce computation.

Phenotype simulation and EBV prediction. We constructed a GRM $\mathbf{G}$ using method 1 of VanRaden (2008), and a pedigree numerator relationship matrix $\mathbf{A}$. For each of 2,000 replicate simulations, we randomly selected a set $\mathcal{C}$ of 1,000 SNPs as causal loci and drew their allele substitution effects independently as $\beta_j \sim \mathcal{N}(0, \sigma_\beta^2)$, with $\sigma_\beta^2 = 1/\left(2\sum_{j\in\mathcal{C}} p_j(1-p_j)\right)$, where p_j is the major allele frequency of SNP j. The true breeding value of individual i was computed as $u_i = \sum_{j\in\mathcal{C}} z_{ij}\beta_j$, where z_{ij} is a genotype code (0/1/2). Training and validation phenotypes were generated as $\mathbf{y}\,[\mathbf{y}_{tr}, \mathbf{y}_{val}] = \mathbf{u}+\mathbf{e}$ with $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$, where $\sigma_e^2 = \text{Var}(\mathbf{u})(1-h^2)/h^2$ for $h^2 \in 0.3, 0.6, 0.9$. To generate two vectors of EBV that were to be compared, the following two models were fitted to phenotypes in training $\mathbf{y}_{tr}$ (Meuwissen *et al*., 2001):

- GBLUP: $\boldsymbol{y} = \mu\mathbf{1}+\mathbf{u}+\mathbf{e}$ with $\mathbf{u}\sim\mathcal{N}(\mathbf{0},\sigma_u^2\mathbf{G})$ giving GEBV $\mathbf{x}_1$ for validation animals
- PBLUP: same model but with $\mathbf{u}\sim\mathcal{N}(\mathbf{0},\sigma_u^2\mathbf{A})$, giving PEBV $\mathbf{x}_2$ for validation animals.

Variance components were estimated by REML based on the full data. Observed PA statistics for each replicate were computed as $r = \text{corr}(\mathbf{y}_{val}, \mathbf{x_1}) - \text{corr}(\mathbf{y}_{val}, \mathbf{x_2})$, $\mathbf{s} = (\check{\mathbf{x}}_1+\check{\mathbf{x}}_2)/\sqrt{2}$, and $\mathbf{d} = (\check{\mathbf{x}}_1-\check{\mathbf{x}}_2)/\sqrt{2}$. To evaluate the GLS-WB test, we held each replicate's validation EBV ($\mathbf{x}_1$, $\mathbf{x}_2$) fixed and generated new validation phenotypes as $\mathbf{y}_{val} = \beta_S\boldsymbol{s} + \beta_D\mathbf{d} + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0}, \mathbf{V})$. Under H_0 we set $\beta_D=0$; under alternative hypotheses, we used $\beta_D \in \{0.04, .., 0.40\}$. For each replicate and each β_D, we computed Δr and obtained a two-sided p-value for the GLS-WB and Williams tests. These p-values were then summarized to estimate T1E under $\beta_D=0$ and size-corrected power under $\beta_D>0$, as described in the following.

Evaluation of T1E and power. For a range of nominal α levels, from 0.01 to 0.20, we estimated empirical T1E as the fraction of the 2000 replicates with $p\leq\alpha$, with uncertainty summarized by 95% Monte Carlo confidence interval. Because the Williams test can be anti-conservative under dependence among elements of $\mathbf{y}_{val}$, nominal-α power curves (i.e. rejecting at observed $p\leq\alpha$) are not directly comparable between the GLS-WG and Williams tests. We

therefore report size-corrected power at a matched empirical T1E rate. For each test and target α (0.05), we first computed a test-specific cutoff τ as the α-quantile of that test's observed p-values under H_0 (such that the empirical T1E under H_0 equals α). Power under the alternative hypothesis was then reported as the rejection rate based on $p \leq \tau$. To reduce runtime, computations were implemented in JAX (Bradbury et al., 2018) on a GeForce 3090 using 2.56 ± 0.04 s for 2,000 replicates, with 1,999 bootstraps per replicate for each h^2 and Δr.

Results

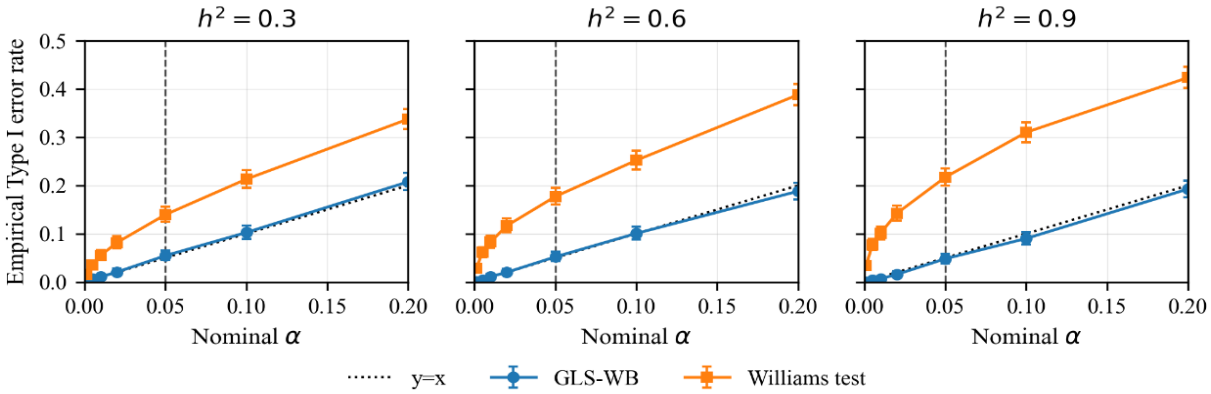

Figure 1. Empirical type I error rates under H_0 of equal accuracies, for different nominal α and heritabilities (h^2). Error bars show 95% confidence intervals. Vertical line marks α=0.05.

Figure 1 shows that, at nominal α=0.05, GLS-WB achieved a T1E of 0.057 (h^2=0.3), 0.052 (h^2=0.6), and 0.048 (h^2=0.9), while the Williams test yielded much higher T1E of 0.140, 0.177, and 0.217, respectively. Fig. 2 reports power of GLS-WB and William's test for a range of true □r, using size-corrected cutoffs τ, such that both methods are compared at the same empirical T1E rate of 0.05. At small Δr (≈ 0.01), both methods achieved power ~0.06-0.08; at larger Δr (≈ 0.10), power rose to ~0.95-0.99. Across all three h^2 evaluated, GLS-WB and the size-corrected Williams test attained similar power. Although a Holm-adjusted McNemar test indicated that the Williams test had statistically higher power at moderate-to-large Δr and higher h^2, differences in power were modest, typically on the order of 1 to 4% points.

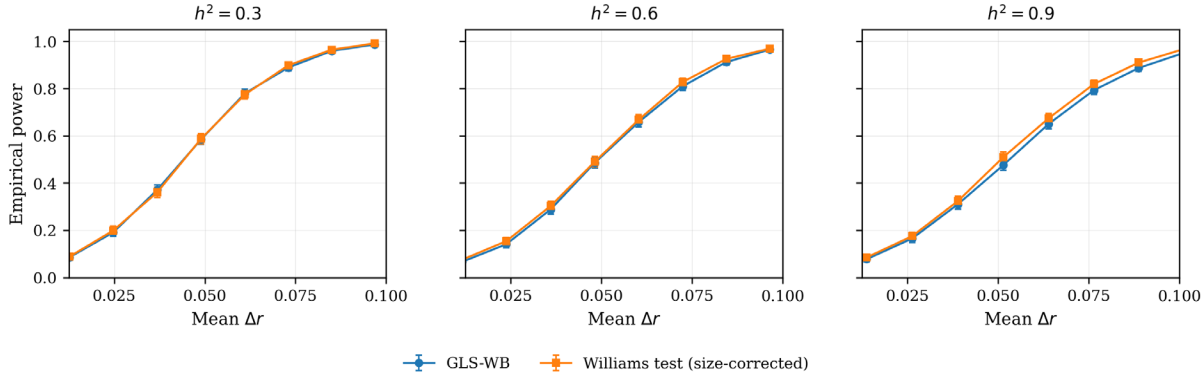

Figure 2. Size-corrected power at nominal α = 0.05, plotted against the realized Δr. Error bars indicate ±1.96 standard errors of power at the size-corrected cutoff τ (empirical τ at H_0: 0.046, 0.047, 0.054 for GLS-WB, and 0.009, 0.004, 0.002 for Williams, at heritability = 0.3, 0.6, 0.9).

Discussion

The Williams test is appealing because it treats the two correlations as paired through their common associated phenotypes, but it implicitly relies on validation records contributing independent information. When validation individuals are related, both the phenotypes and the two EBV vectors inherit a dependence structure induced by shared genetic background (and

shared connections to the training data). This reduces the effective sample size and increases the sampling variability of the PA correlations. Because the Williams test does not account for this dependence, it can underestimate the standard error of PA comparisons and rejects the null hypothesis of no difference too often, as seen in Fig. 1. The inflation becomes more severe as h^2 increases because of larger correlations between relatives. Our proposed GLS-WB addresses this dependence by framing the comparison as a paired inference problem of equality of PA correlations for two sets of EBV conditional on the realized EBV vectors. By decorrelating the validation phenotypes using a working covariance model and applying a wild bootstrap that preserves the paired structure, GLS-WB maintains the targeted T1E rate based on the set nominal rate under relatedness of validation individuals, as illustrated in Fig. 1, without changing the practical workflow of computing PA. This also implies that GLS-WB can be used to compute statistical power of accuracy comparisons without size correction (Fig. 2). This is in contrast to the Williams test, which resulted in severely inflated TIE rates and required size correction to obtain relevant power estimates. After such size correction, power of GLS-WB was very similar to that of the Williams test (Fig. 2).

The GLS-WB does, however, require a reasonable working covariance matrix for the validation phenotypes (and thus depends on assumptions about genetic relationships and variance components), and its operating characteristics can be sensitive to misspecification of h^2 and other sources of heterogeneity. In addition, the simulation that was used here to evaluate GLS-WB simplified the training–testing phenotype covariance, which resulted in the mean of the test statistics to be zero, which would not be true in real data. Methodological extensions that address this issue are being developed in subsequent work. Conceptually, GLS-WB can be formulated under any linear mixed model for the validation phenotypes by augmenting the current additive genomic model with additional dominance or epistatic effects. As with any validation analysis, leakage between training and validation can distort inference and should be avoided, e.g., when close relatives are shared across sets or preprocessing uses validation data. While the developed GLS-WB was based on the validation data consisting of individual phenotypes, extensions to derived quantities such as deregressed EBV are plausible but require explicit handling of the reliability-dependent heteroscedastic error. Extending the GLS-WB test to LR validation statistics (Legarra and Reverter, 2018) requires explicit modelling of the joint sampling covariance of the partial and whole EBV vectors.

Conclusions

Our proposed GLS-WB provides a dependence-aware test for comparing validation accuracies of genetic predictions in related cohorts. Using simulations, GLS-WB was shown to control T1E, while the Williams test can be severely anti-conservative. When compared at equal empirical T1E rates, GLS-WB and Williams test achieved comparable power for a range of accuracy differences. The GLS-WB test can be used whenever an appropriate working covariance for the validation response is available.

References

Cleveland, M. A., Hickey, J. M., and Forni, S. (2012) G3 2: 429-435. doi:10.1534/g3.111.001453.
Bradbury, J., Frostig, R., et al. (2018) JAX (version 0.3.13). GitHub. URL: http://github.com/jax-ml/jax.
Davidson, R., and Flachaire, E. (2008) J. Econometrics 146: 162-169. doi:10.1016/j.jeconom.2008.08.003.
Legarra, A., and Reverter, A. (2018) Genet. Sel. Evol. 50: 53. doi:10.1186/s12711-018-0426-6.
Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001) Genetics 157:1819 doi:10.1093/genetics/157.4.1819
Momin, M. M., Lee, S., Wray, N. R., et al. (2023) Am. J. Hum. Genet., 110, 349 doi:10.1016/j.ajhg.2023.01.004
Steiger, J. H. (1980) Psychological Bull. 87: 245-251. doi:10.1037/0033-2909.87.2.245.
VanRaden, P. M. (2008) J. Dairy Sci. 91: 4414-4423. doi:10.3168/jds.2007-0980.
Williams, E. J. (1959) J. Royal Stat. Soc., Series B 21: 396-399. doi:10.1111/j.2517-6161.1959.tb00346.x.