Updated Interbull software for genomic validation tests

P.G. Sullivan
Lactanet, 660 Speedvale Ave, Guelph, Ontario, Canada, N1K 1E5

Abstract

The Interbull GEBV test software was updated by adding new options and modernized methods for the validation of genomic predictions being used currently to select and market dairy bulls internationally. Misaligned scales of evaluation can now be detected by the program, and the scales of full versus reduced data properly aligned before conducting validation tests, to reduce bias in the validation test results. Genomic pre-selection bias can also be reduced by using new validation target options added to the program. The new methods were applied to Canadian genomic evaluations with full data from August 2022, and reduced data equivalent to August 2018, for 30 MACE traits evaluated in Canada for each of the Holstein (HOL), Jersey (JER), and Ayrshire (RDC) breeds. Evaluation scale alignments were needed for these data and reduced the bias of validation test slopes by as much as 10% for some traits. Validation slopes were generally closer to 1 for all traits and breeds when using genomic validation targets that do not include GPS bias. However, the smaller deviations from desired regressions close to 1.0 were partly due to higher auto-correlations with GEBV in reduced data, when using a genomic validation target instead of the current target that does not include genomic information. New output files can be used to isolate data or modelling issues that might be causing poor validation test results, and thereby facilitate modelling improvements. A new suite of validation tests was made possible by the software updates, which could be based on tests for bias in the average evaluations of animal groups.

Key words: genomic validation, software, GEBV, GMACE

Introduction

Validation of genomic evaluation systems is important for quality assurance of the national GEBV used to market young dairy bulls, and as input for GMACE international evaluations published by Interbull. The official GEBV validation test of Interbull (Mäntysaari et al, 2010) was developed during the early years of genomic selection, when national EBV computed without genotypes could still be considered unbiased for most recently proven AI bulls. The new AI bulls at that time had been randomly selected for use in AI prior to the genomics era.

After many years of genomic selection, the EBV of most recent AI bulls can no longer be considered unbiased, because these newer bulls were genomically pre-selected (GPS), and the effects of GPS are ignored when genotypes are excluded to compute the national EBV. The EBV are said to include a GPS bias because the effects of GPS were not properly included in the national EBV predictions.

Validation tests for the national EBV of recent bulls used as input to MACE are also under review, due to concerns that GPS biases in national EBV will adversely affect MACE results. The current trend validation tests of input EBV used in MACE (Boichard et al, 1995) are also losing power, because bulls are now being replaced at much younger ages (Mäntysaari and Kudinov, 2022).

Purposes of the present study were to update software for the Interbull GEBV validation tests, and to develop and gain practical experiences with new validation methods for genomic evaluation systems, which are more applicable to the current data after many years of international GPS in dairy cattle populations. Since new validation tests are needed for both the GEBV and EBV used as input to Interbull international evaluations, the potential expansion of GEBV test software for the purposes of validating both GEBV and EBV is also of interest. Interbull continues to provide both genomic and non-genomic international

evaluation services, through GMACE, Intergenomics, SNP-MACE, MACE, and truncated-MACE services.

Materials and Methods

Interbull software used currently for the GEBV tests (gebvtest.py) was updated by adding several new features and options. New validation tests were then applied to Canadian genomic evaluations from full data in August 2022, and reduced data equivalent to August 2018, for 30 MACE traits evaluated in Canada for each of the Holstein (HOL), Jersey (JER), and Ayrshire (RDC) breeds.

1 - Interbull validation software updates

The new features and options facilitate research and development of new validation methods that consider the following:

1. Publication scales that are data-dependent can differ between predictor and predictand in validation models, causing bias in validation test results.
2. The data available for validation tests has changed, reducing the power of current tests because AI bulls are now replaced at younger ages due to genomic selection.
3. While modernizing GEBV validation tests, EBV tests could also be updated to align with the new GEBV tests.
4. GPS biases are being ignored in the current Interbull validation tests, and changes should be made to account for GPS.
5. Auto-correlations due to same information in predictor and predictand can reduce testing power and complicate the interpretation of validation test results.
6. New information to help isolate problems can facilitate model improvements.

Regarding #1, methods were developed to minimize bias in validation test results that are caused by scale differences. A change in the evaluation scale from reduced to full data can now be addressed with a new option to detect and correct the difference, aligning the scales before conducting validation tests. If scales are the same and alignment is not needed, the changes made by the software are trivial and do not impact the validation test results, so the option can be used without consequence with such data. When scales are misaligned, it is important for test bias reduction to always use:

--baseadj={EBV, GEBV}

Many new validation tests can be added after aligning the evaluation scales. A non-zero average change for evaluations aligned to the same scale can be treated as evidence of bias in reduced-data evaluations for any group of individuals. New checks for bias can therefore be added for a battery of new tests targeting specific groups of interest, such as average bulls, top bulls, local bulls, imported bulls, cows, etc. Considering #2, new tests will likely be needed to address the evolution of data now available for validation purposes, and a growing obsolescence of current tests (Mäntysaari and Kudinov, 2022).

In the Interbull trend test III, a nuisance factor is included in the validation model to account for scale differences between data sets. This nuisance factor would no longer be needed after aligning the scales with --basadj. Considering #3, a simpler validation model could be developed and added as a future update in gebvtest.py, to replace separate programs used currently to apply trend test III.

To explore #4 and #5, which are related, new options allow the use of different validation targets, and using either weighted (WLS) or unweighted (OLS) least-squares regression tests. The current Interbull tests use WLS, while a growing number of validation test results in the literature have been based on OLS, following LR regression tests proposed by Legarra and Reverter (2018, 2019):

--target={DEBV, DGEBV, DGPA, VFEBV, EBV, GEBV}
--weight={ITB, LR}
--min_byear={yyyy}

The following combination of options is equivalent to LR regression using OLS: --baseadj=GEBV, --target=GEBV, --weight=LR. Additionally, the --min_byear option can be used to modify how many bulls are included in the LR regressions. The addition of older bulls with progeny in both reduced and full data sets will increase auto-correlations in OLS, but not in WLS regressions used for Interbull validation

tests, because Interbull uses weights of zero for the older progeny-proven bulls.

For #6, new features were added to build on the --mergefiles option already available in the current program, which is used to create output data files for all traits being tested. These trait data files are provided for input to post-analysis programs written by the user, for the purpose of isolating evaluation system problems and developing better model(s) for the problematic trait(s). Some new options were added to make it easier to focus on only a subset of the traits:

--traitsincl={tr1,tr2,…}
--outdir={path}
--baseincl={min,max byr : proof_types csv_list : proof_status csv_list : official Y/N}

The --traitsincl option limits application of the validation tests, and along with that the creation of output data files for post-analyses, to only the specified trait(s) of interest. The --outdir option makes it easier to use separate locations, if desired, for the data and summary files created when using different combinations of options.

The --baseincl option allows users to check if scale alignments are affected by using fewer or more bulls in --baseadj, which can increase confidence in the generality of the approach. If necessary, the scale alignments can also be optimized differently for each trait or group of traits, if expanding or restricting the sets of bulls used in --baseadj is appropriate for different data scenarios.

The file300 fixed length format was updated to allow either fixed or variable length records, and any combination of white-space and commas to delimit the fields. Additional fields can also be appended after the required fields, and the extra data will be included in output trait data files created with --mergefiles, thus making it available for modelling in the user-developed post-analysis programs, or for other investigative purposes. Extra fields of interest could include alternative identifiers for animals, animal details such as gender, and any variables of interest for investigative validation models designed to isolate weak areas of a genetic evaluation system (VanRaden, 2021).

2 - Accounting for different evaluation scales

Without exception, all genetic (EBV) and genomic (GEBV) evaluations are expressed on a scale that is relative to an arbitrarily defined genetic base. Differences between data sets that alter either the animals included in the genetic base or the statistical properties of evaluations for a same set of base animals, can have a direct impact on estimated parameters for a validation test. If evaluations are published as solutions obtained directly from linear mixed model equations (MME), without additional base adjustments, then the evaluation scale is defined by intrinsic constraints on random effects in linear mixed models, for example that $1'A^{-1}\hat{u} = 0$. In most genetic evaluation models, the MME constraint forces average evaluations to be zero for the group of individuals with unknown parents, or for genetic groups of unknown parents at the top of the pedigree, which define the implied genetic base. The evaluation scale is thus relative to base animals or genetic groups. Changes in pedigree edits or genetic group definitions, or the addition of new animals with unknown parents, can alter the base definition and thus the evaluation scale between data sets. These differences would normally be small, but it is also unusual for dairy cattle genetic evaluations to be reported this way, without any additional base adjustments.

For dairy sires, genetic evaluations are commonly expressed relative to a pre-defined group of bulls or cows born within a specified range of years, and usually with some minimum degree of reliability for their evaluations. For example, the base group could be a set of recently proven bulls, or a group of cows with performance data included in the evaluation. The evaluation scale is shifted in mean by subtracting the average solution of base group individuals, and adding a preferred base group average, which can be any arbitrary value such as 0, 10 or 100. The variance of published evaluations can also be scaled up or down, so evaluations of base group individuals have a pre-defined standard deviation (SD), such as 5 or 10. In Canada, nearly all genetic evaluations for dairy cattle are base-adjusted to standardize both mean and SD of base group individuals, and with base groups updated annually the means and SD of base individuals can change significantly between full versus reduced data,

prior to standardization. After standardization, the genetic evaluation scales are no longer directly comparable if different standardization factors were used for the two data sets.

It is assumed in Interbull GEBV tests that scales of evaluation are the same and that full and reduced data evaluations are therefore directly comparable. This would rarely be true for Canadian dairy cattle evaluations, nor for other countries publishing EBV and GEBV as relative breeding values, standardized for both mean and SD of a base group. There are at least four ways to minimize differences between evaluation scales of the Y and X regression variables used in validation tests, where Y are from full data, and X from reduced data:

1. Use the same group of base animals to adjust the evaluation scales of Y and X. Note that average reliabilities of base animals might differ between Y and X data.
2. Use different base groups to ensure similar reliabilities between Y and X. Note that true genetic variances might differ between the different base groups of individuals.
3. Use a single set of base adjustments for both Y and X, instead of using different ones.
4. Compare a subset of evaluations, for individuals with identical expectations for mean and SD in both Y and X, to estimate and correct for any scale misalignment.

Options one through three might be sufficient if evaluation scales are not variance-standardized, but for any scenario where the scales for X and Y are variance-standardized, where phenotypic data are pre-adjusted or adjusted within the model for heterogeneous variances (HV), or when data transformations are used (e.g. for categorical data), then option four should be preferred.

With option one, the increasing EBV variance in full data, which is due to higher reliabilities, gets removed by standardizing EBV variance to be the same in both data sets, thus shrinking the SD of full data evaluation scales relative to the reduced. With option two, the SD of evaluations from full data could be shrunken or expanded relative to the reduced, if changing goals or accuracies of selection have had different effects on genetic variance, with differently selected animals included in the different base groups. Although it is generally preferred over one and two for the purpose of validation testing, option three might be difficult to implement in practice, depending on the extent of required program modifications. The program changes for validation testing are not likely to be needed for any other purpose, and they might also ignore potential scale differences due to HV adjustments or data transformations being applied differently between the two data sets.

Option four is a general approach that does not require modifications to existing evaluation systems, knowledge of data adjustments or transformations that might have been used, or details about the base group definitions or evaluation properties of base group individuals. The idea behind option four is that BLUP predictions should only change if newly added data are sufficient to increase reliabilities, and if the mean or variance of evaluations changes for individuals with no changes in reliability, then the scale of expression must have changed between the two data sets. The observed distribution changes for these individuals can therefore be used to estimate a regression equation that aligns the two evaluation scales, thereby ensuring evaluations are directly comparable between the two data sets.

The expected changes for BLUP predictions after adding new data have a well-defined and simple distribution (Klei et al, 2002; Legarra and Reverter, 2018, 2019), which can be expressed as:

$$V(\delta\hat{u}) = \delta\hat{R} * V(u)$$

The $\delta\hat{u}$ and $\delta\hat{R}$ denote changes in genetic predictions and reliabilities respectively, after adding new data, and $V(u)$ is the true genetic variance. If $\delta\hat{R} = 0$, then the expected $V(\delta\hat{u}) = 0$ and the regression of genetic predictions from full data ($\hat{u}_{full}$) on reduced ($\hat{u}_{red}$) should have an intercept of 0, a slope of 1, and a model R^2 very close to 1. Deviations from the unity regression line reflect a change in the evaluation scale.

Approximate reliabilities are adequate to identify progeny-proven bulls in reduced data with relatively small increases in information in the full data. From the full and reduced evaluations of these bulls, the following

prediction equation is estimated and then used to correct the scale of reduced-data evaluations:

$$\hat{u}_{full} = a_{base} + b_{base} * \hat{u}_{red} + e$$

Predictions from this equation are evaluations from reduced data, re-scaled to the evaluation scale of full data. This re-scaling makes evaluations from reduced data directly comparable to the full-data evaluations.

A weighted regression helps to control standard errors (SE) of prediction for $\hat{a}_{base}$ and $\hat{b}_{base}$, by keeping the SE relatively low while still heavily weighting the evaluations of bulls with lowest $V(\delta\hat{u})$. The regression weights are defined as:

$$W_{base} = \hat{R}_{red} * c^{\delta\hat{R}},$$
$$\hat{R}_i \in [0,100]$$
$$c \in [0.25,0.75]$$

where $\hat{R}_{red}$ is the approximated reliability from reduced data, and subscript i in $\hat{R}_i$ refers to either full or reduced data. The power function used for these weights decreases exponentially with higher $\delta\hat{R}$, and most quickly when using a smaller base value c. The smallest values of c can be used when many bulls have $\delta\hat{R} = 0$, as is usually the case with --baseadj=EBV, because there are many historical progeny-proven bulls no longer in service and adding no new daughter information in the most recent four years. It is possible with --baseadj=GEBV, however, to have few bulls with $\delta\hat{R} = 0$, especially for newly recorded traits with notable population increases in genomic reliability from the four years of most recent phenotypic data.

To accommodate all data of potential interest, an optimization was therefore included in a generalized implementation of --baseadj. A balance was needed between maximizing the proportional weight on data with $\delta\hat{R} = 0$ (using smallest c=0.25), versus reducing the SE of $\hat{a}_{base}$ and $\hat{b}_{base}$ by spreading regression weights across more bulls, who would still have $\delta\hat{R}$ close to zero, but not exactly equal to zero (using c>0.25). A Newton-Raphson (N-R) iteration is used to quickly locate optimal power base values c for each data set. The solved equation, based on evaluations of n progeny-proven bulls, is $\frac{1}{n}\sum c^{\delta\hat{R}} = 0.04$, which creates a sum of weights in WLS that is 4% of the equivalent OLS weighting w.r.t. $\delta\hat{R}$, where OLS is analogous to assuming $\delta\hat{R} = 0$ for all bulls (c^0=1 for any c). Effectively, only 4% of the available data on proven bulls is then used to align the scales, while weighting most heavily the bulls with smallest $\delta\hat{R}$. Any values of c outside the range 0.25-0.75 are replaced by the range limits.

The choice of 4% as an optimum proportion of data for scale alignment was based on a previous study comparing relative variances of estimated regression coefficients across 37 MACE traits, for scale alignments versus validation regression tests. The Canadian evaluations published in 2018 were used to create reduced data equivalent to 2014 in that study, where it was feasible to use c=0.25 for all traits with --baseadj=EBV but optimal values for c were higher and more variable across traits (c=0.53±0.14) with --baseadj=GEBV. Many more years of historical EBV than GEBV were included in that study, and the higher optimal values estimated for c with GEBV were due to including only the most recent years of GEBV. Despite very different sets of bulls with GEBV versus EBV, the estimates for $\hat{a}_{base}$ and $\hat{b}_{base}$ were still very similar across all traits, due to the optimizations of c across traits when using GEBV. In more recent tests, we included the GEBV of more historical bulls, and c=0.25 can now be used for all traits with either EBV or GEBV in --baseadj. The estimated scale-alignment equations were almost identical between --baseadj=EBV and --baseadj=GEBV across all traits of HOL, JER and RDC in the present study.

3 - Improving the validation target

A potential improvement to the GEBV test would be using an unbiased validation target that includes proper estimates of GPS effects for the GPS bulls. This can be achieved by using GEBV as the target variable (e.g. Legarra and Reverter, 2018, 2019), or by adding GPS effects into the EBV, by generating EBV with a model that includes GPS-effects while not requiring genotypes (Sullivan et al, 2023). A de-regression of GEBV can also be used to expand the variance and create a d-GEBV validation target that is more like phenotypic daughter averages (VanRaden, 2021). The d-GEBV is a logical genomic alternative to d-EBV used in

the current test. As with d-EBV currently, the d-GEBV can be weighted for highest emphasis on the bulls adding relatively more information for cross-validation in the recent data.

A weighted regression can lessen concerns that use of a genomic validation target increases auto-correlations between the predictor (X) and predictand (Y) in cross-validation tests. Auto-correlations are due to the common information between X and Y. When both X and Y are GEBV, the estimated sums of SNP effects are included on both sides of the validation regression equation, and this shared information is substantially more than with the current test that shares only a common estimate of the parent average (PA), with X being a GEBV and Y a d-EBV.

Results and Discussion

1 - Using the updated software

The modified software can be run the same way as before, and while using the same input files as before, if not using the new options. To take advantage of new options, changes might be needed in the input files provided to the program. The new requirements are detailed in a document provided with the software, but in general terms:

- New input file(s) are needed to --target the new genomic alternatives to d-EBV.
- Adding records for more animals to the current input files might improve scale alignments when using --baseadj.
- A new, optional input file allows users to provide their own customized validation target, so the software can be used for internal tests of interest in addition to validation tests required by Interbull.
- Users should write their own programs for post-analyses, to isolate problems specific to their evaluation systems and data. New "BaseCorrected" output files are now available to help with those efforts, in which the reduced evaluations have all been re-scaled by --baseadj, to be expressed on the same scale that was used for full data.

2 - Accounting for different evaluation scales

To align evaluation scales between the reduced and full data sets of all breeds and traits, the new software option --baseadj=EBV was used. For Canadian data, the percentages of regression weights on EBV of bulls with zero (integer-rounded) changes in reliability ranged from 97.1-99.9% across 36 traits in HOL, 98.0-99.9% across 32 RDC traits, and 96.0-99.9% across 30 JER traits. The regression model R^2 were higher than 0.98 for HOL, 0.97 for JER, and 0.96 for RDC, and the estimated --baseadj slopes ranged from 0.90-1.10 for HOL and RDC, and 0.91-1.10 for JER.

The estimated slopes to align evaluation scales for several conformation traits are shown in Table 1. For all three breeds, estimated slopes for scale alignment were consistent with ratios of SD standardization factors used for the two data sets. Scale differences created artificially by using different standardization factors were effectively reverted with --baseadj.

Table 1. Ratio of SD standardization factors used in reduced versus full data (SD-Adjust), versus --baseadj estimated slopes to align the evaluation scales ($\hat{b}_{base}$).

	SD-Adjust : $\hat{b}_{base}$ (*100)		
Trait	HOL	JER	RDC
STA	94 : 91	107 : 105	104 : 104
CWI	99 : 98	100 : 100	99 : 97
BDE	94 : 93	98 : 98	100 : 99
RWI	97: 96	101 : 99	97 : 98
RAN	100 : 99	103 : 101	104 : 103
FAN	108 : 107	111 : 110	100 : 98
RLS	97 : 96	103 : 102	107 : 106
UDE	93 : 92	103 : 102	106 : 107
USU	103 : 101	103 : 102	100 : 99
FUA	92 : 90	106 : 105	99 : 98
FTP	102 : 101	103 : 101	110 : 110
FTL	92 : 92	93 : 91	97 : 97
RUH	98 : 93	99 : 98	98 : 98
RTP	103 : 102	102 : 99	96 : 95

The trait stature in Holsteins required one of the largest adjustments for scale alignment ($\hat{b}_{base}$=0.91). The SD of breeding values for stature has increased significantly in recent years, due to an increasing focus on feed efficiency traits, a greater interest in moderate cow size to control maintenance feed costs, and a wider range of views regarding selection for stature. The within-year SD for stature of proven bulls has increased and was much higher in the four most recent years included in the base group for full data only, compared with the oldest four years included in the base group for

reduced data only. The increased variability for stature of base bulls, after rolling the base group forward by four years, was removed when both evaluation scales were adjusted to force the same SD=5 on the EBV of these very different base groups between the two data sets.

Aligning evaluation scales with --baseadj guarantees a consistent use of the well-defined contrast between Validation versus Base Adjust bulls, as the basis for validation tests. For Holstein stature, Figure 1 shows the distribution of genomic reliabilities in full versus reduced data for these two groups of bulls. Validation bulls have large $\delta\hat{R}$, while the Base Adjust bulls have small $\delta\hat{R}$.

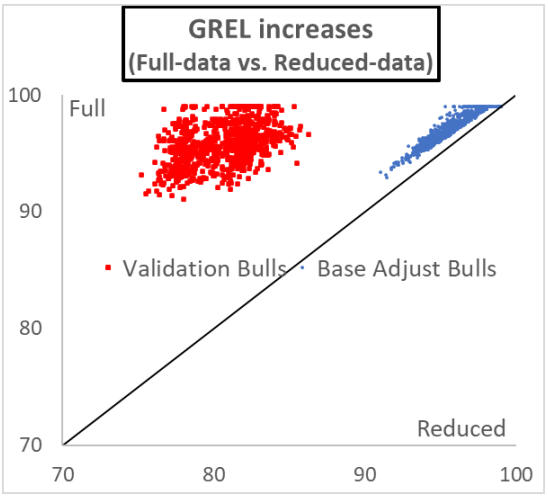

Figure 1. Genomic reliabilities (GREL) from full versus reduced data for Holstein stature, for the Base Adjust bulls used to align evaluation scales, and for validation bulls used to test for bias in reduced-data GEBV.

In Figures 2 and 3, simple regressions for Validation and Base Adjust bulls are shown in the top left and bottom right quadrants respectively, with data and regressions before scale alignment in Figure 2 and afterwards in Figure 3. These simple regression estimates differed only slightly from weighted regressions computed and used within the software. The estimated Base Adjust slope in Figure 2 ($\hat{b}_{base}$=0.9078) indicates that SD for evaluation scales of stature were arbitrarily reduced by almost 10% more in the full versus reduced data, after rolling the base forward four years.

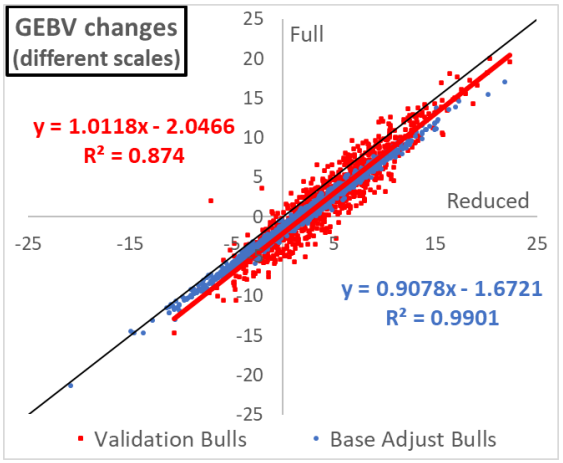

Figure 2. Full data GEBV (y) versus reduced data GEBV (x) for Holstein stature, before aligning the two evaluation scales.

After aligning the evaluation scales, regression lines for both the Base Adjust and Validation bulls passed through or very close to the origin, and with the slope for GEBV of Base Adjust bulls being essentially equal to 1.00 (Figure 3). The very small difference from a unity regression for Base Adjust bulls was due to small differences between simple versus weighted regressions, and additionally between regressions of GEBV in Figure 3 versus EBV within the program with --baseadj=EBV.

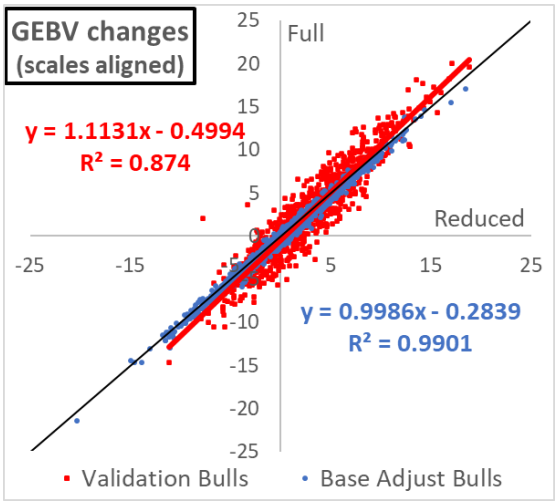

Figure 3. Full data GEBV (y) versus reduced data GEBV (x) for Holstein stature, after aligning the two evaluation scales.

Before aligning the evaluation scales for Holstein stature, a validation slope of 1.012 (Figure 2) was estimated from misaligned EBV that were not directly comparable. This was a

very biased validation test result. The validation slope with --baseadj=EBV was much higher at 1.113 (Figure 3). The ratio of slopes from simple regressions for Validation versus Base Adjust bulls was identical both before and after aligning the scales (i.e. 1.0118/0.9078 = 1.1131/0.9986 = 1.115 from Figures 2 and 3 respectively). In both cases, this ratio matched the unbiased, weighted validation test slope of 1.113 with --baseadj=EBV. The interpretation of results is therefore the same before as after aligning the scales, if we consider the properties of GEBV for new young bulls adding daughters relative to highly proven bulls with very stable evaluations, because they are no longer used or because they already had reliabilities close to 100% in the reduced data. When interpreting validation results in this relative way, we always conclude that GEBV of young bulls were under-dispersed by 11% for this example trait. The test result was worse after --baseadj than before because the misalignment of scales was hiding a dispersion problem.

For all three breeds, the correlations between estimated slopes for scale alignment versus validation tests were close to zero (-0.24, 0.14 and -.33 for HOL, RDC and JER respectively), confirming that --baseadj did not systematically or inappropriately make it easier or harder to pass the validation tests across the traits. The scale alignments were needed for Canadian evaluations, to remove bias in the validation test results because evaluations for most traits are expressed as relative breeding values in Canada, after scaling for variance differently in reduced versus full data.

3 - Improving the validation target

The current GEBV test of Interbull uses d-EBV as the validation target. Results using the current test are summarized in Table 2, for MACE traits evaluated in Canada for all three breeds included in the present study. Validation test slopes were close to 1.00 for production, SCS and conformation traits of all breeds, with acceptably small practical differences from 1.00 for HOL, and with statistically small differences from 1.00 for JER and RDC, these latter breeds having much higher SE compared to HOL. The number of validation test bulls for the three groups of traits respectively were: 856, 877 and 832 for HOL; 69, 69 and 59 for JER; and 52, 54 and 43 for RDC.

The validation test slopes were much lower than 1.00 for the other traits of longevity, fertility, and workability. Genomic evaluation methods used for these other traits need to be reviewed, considering the poor validation test results across all three breeds.

Table 2. Average ±SD of validation slopes ($\hat{b}$), with de-regressed EBV as the validation target, and average SE of $\hat{b}$ in parentheses.

Trait Groups	HOL	JER	RDC
Production & SCS (4 traits)	1.05 ± .03 (0.02)	0.92 ± .15 (0.09)	0.74 ±.08 (0.15)
Conformation (19 traits)	1.00 ± .10 (0.03)	0.81 ± .16 (0.12)	0.90 ± .32 (0.18)
Other (7 traits)	0.66 ± .11 (0.05)	0.26 ± .25 (0.16)	0.69 ± .29 (0.23)

Using d-EBV as the validation target has been criticized because of the GPS bias issue with EBV. Comparative validation test results were generated using d-GEBV (Table 3) and GEBV (Table 4) as genomic alternatives to the d-EBV used in current official tests of Interbull. Using the genomic validation targets can reduce concerns about GPS bias impacting test results. With d-GEBV, the average validation slopes were generally closer to 1.00 across all breeds and trait groups, the SD and SE were generally lower for HOL, and higher for JER and RDC (Tables 3 versus 2).

Table 3. Average ±SD of validation slopes ($\hat{b}$), with de-regressed GEBV as validation target, and average SE of $\hat{b}$ in parentheses.

Trait Groups	HOL	JER	RDC
Production & SCS (4 traits)	1.05 ± .02 (0.02)	1.00 ± .18 (0.09)	0.80 ±.09 (0.15)
Conformation (19 traits)	1.02 ± .08 (0.03)	0.99 ± .16 (0.13)	0.95 ± .35 (0.20)
Other (7 traits)	0.71 ± .11 (0.03)	0.46 ± .41 (0.22)	0.64 ± .53 (0.33)

The average validation slopes moved even closer to 1.00, and the SD and SE were reduced for all traits and breeds, when using GEBV as

the validation target instead of d-GEBV (Tables 4 versus 3).

Table 4. Average ±SD of validation slopes ($\hat{b}$), with GEBV as the validation target, and average SE of $\hat{b}$ in parentheses.

Trait Groups	HOL	JER	RDC
Production & SCS (4 traits)	1.05 ± .02 (0.02)	1.00 ± .14 (0.08)	0.82 ±.08 (0.13)
Conformation (19 traits)	1.01 ± .07 (0.02)	0.97 ± .13 (0.10)	0.97 ± .29 (0.17)
Other (7 traits)	0.76 ± .10 (0.02)	0.56 ± .32 (0.14)	0.84 ± .28 (0.21)

Because the SD and SE of validation slopes were generally reduced in tandem, when changing from d-EBV to d-GEBV to GEBV, the results of Interbull statistical tests that require $|t|<2$ were similar for all three of these validation targets. The main impact of changing validation targets for the Canadian data was therefore on the probability of exceeding practical limits of tolerance used in the current official Interbull test. The lower and upper limits of tolerance are currently the expected value of $\hat{b}$-0.10 and 1.20 respectively. These limits should be reviewed before defining a new official test for Interbull that will be based on a new validation target that changes the SD of test results in practice.

The importance of reviewing practical limits for official tests, which can have important impacts on test failure rates, is further emphasized by comparing validation test slopes ($\hat{b}$) using WLS in Interbull regression tests versus OLS in LR regression tests, when including the same validation bulls born since 2014 in both, and after additionally including older progeny-proven bulls in the LR regressions. The maximum values for $|\hat{b}-1|$ decreased when changing the WLS validation target from d-EBV to d-GEBV to GEBV (Table 5), consistent with corresponding reductions in SD of $\hat{b}$ described above. Results were nearly the same for WLS versus OLS when the LR regressions included the same bulls born only since 2014, but they decreased significantly when including additional progeny-proven bulls born in earlier years in the LR regressions.

Table 5. Maximum absolute difference from unity ($|\hat{b}$-1|) across 19 Conformation traits, with different validation targets in weighted Interbull regressions of proven bulls born since 2014, and with LR regressions of GEBV including the same and older bulls.

Target	HOL	JER	RDC
d-EBV	0.23	0.36	0.63
d-GEBV	0.16	0.36	0.77
GEBV	0.15	0.32	0.58
LR 2014[z]	0.14	0.31	0.60
LR 2010	0.15	0.16	0.23
LR 2005	0.12	0.12	0.12
LR 2000	0.09	0.11	0.08

[Z]Minimum birth year of proven bulls included in the LR regression GEBV test.

The observed changes in variance and range of validation test slopes are very highly related to increased auto-correlations between the regression variables Y and X, as we move from top to bottom of Table 5. The trend of increasing degrees of auto-correlation is clearly demonstrated by patterns of increasing R^2 for the regression models, moving from top to bottom of the corresponding Table 6. Caution is recommended when interpreting and comparing LR regressions, Interbull GEBV test regressions based on different validation targets, or other regression estimates that might be found in the literature and used for the purpose of genomic validations.

Table 6. Average R^2 of validation regressions across 19 Conformation traits, with different validation targets in weighted Interbull regressions of proven bulls born since 2014, and with LR regressions of GEBV including the same and older bulls.

Target	HOL	JER	RDC
d-EBV	0.51	0.39	0.30
d-GEBV	0.65	0.45	0.30
GEBV	0.69	0.56	0.39
LR 2014[z]	0.70	0.57	0.39
LR 2010	0.85	0.77	0.70
LR 2005	0.91	0.84	0.85
LR 2000	0.94	0.86	0.89

[Z]Minimum birth year of proven bulls included in the LR regression GEBV test.

The current limits of tolerance remain unchanged in the updated software, because a review of best limits for an updated official test has not been completed, and because a new

official validation target has not been finalized. Changes in test failure rates observed with different validation targets should therefore be disregarded while testing the current updated version (gebvtest_2023A.py).

Summary and Conclusions

Evaluation scale alignments will often be needed to avoid bias in validation test results, which were as high as 10% for some traits in recent tests with Canadian data. A new feature was therefore added to the software to properly align evaluation scales of reduced versus full data before applying the validation tests. The current validation target should be changed to a new variable that is not biased by unaccounted GPS effects, while also considering that the variability of observed validation regression coefficients will be smaller if auto-correlations between predictand and predictor are increased by the change of validation target. The updated software can be used to develop and test many new options for GEBV validations, which could be based on average evaluation changes for groups of individuals. The Interbull EBV validation tests are also in need of updating, and development of new tests to replace current EBV validation tests is ongoing.

Acknowledgements

Helpful discussions within the Interbull validations working group are gratefully acknowledged, with contributions from Esa Mäntysaari, Paul VanRaden, Zengting Liu, Raphael Mrode, and Valentina Palucci.

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. J. Dairy Sci. 78, 431-437.

Klei, B., Mark, T. Fikse, F. and Lawlor, T. 2002. A method for verifying genetic evaluation results. Interbull bulletin 29:178-182.

Legarra, A. and Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Gen. Sel. Evol. 50:53. https://doi.org/10.1186/s12711-018-0426-6 .

Legarra, A. and Reverter, A. 2019. Correction to: Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. Gen. Sel. Evol. 51:69. https://doi.org/10.1186/s12711-019-0511-5.

Mäntysaari, E.A., Liu, Z. & VanRaden, P.M. 2010. Interbull validation test for genomic evaluations. Interbull bulletin 41, 17-21.

Mäntysaari, E.A. & Kudinov, A.A. 2022. Usability of different genetic evaluation validation tests in a population subjected to a strong genomic selection and in testing the single-step genomic evaluations. Interbull bulletin 57:106-110.

Sullivan, P.G., Mäntysaari, E.A. & de Jong, G. 2023. Implementation of GPS-MACE accounts for genomic preselection. Interbull bulletin 58…

VanRaden, P.M. 2021. Improved genomic validation including extra regressions. Interbull bulletin 56: 65-69.