# Validation of National Genomic Evaluations

***M.A. Nilforooshan[2], B. Zumbach[1], J. Jakobsen[1], A. Loberg[1], H. Jorjani[1] and J. Dürr[1]***
*[1]Interbull Centre, [2]Department of Animal Breeding and Genetics*
*Swedish University of Agricultural Sciences, Uppsala, Sweden*
*Mohammad.Nilforooshan@slu.se*

## Abstract

Several countries enrolled in Interbull services have developed national genomic models for computation of genomic breeding values. For international trade, these models should be validated. A new validation method was developed and tested on 13 different populations for protein yield. This new method is described in this paper. Because of selective genotyping of young bulls, the slopes of the linear regressions and their expected values differed from one. Comparing these slopes to their expected values, 11 of the 13 populations passed the genomic validation test successfully. Using information on direct genomic values in addition to parent averages doubled the accuracy of predicting future breeding values.

## Introduction

Validation of national genomic evaluations, e.g. by the method proposed by Mäntysaari *et al.* (2010) is a prerequisite for any form of international genomic evaluation, e.g. GMACE (Sullivan and VanRaden, 2009). Nowadays, genomic breeding values (GEBV) are used to select young bulls, because GEBVs are expected to predict proven EBVs more accurately than parent averages alone (Kistemaker and Sullivan, 2010). The goal of the Interbull genomic validation method is to test the unbiasedness of national genomic evaluations by testing how correctly future EBVs can be predicted from current GEBVs (Mäntysaari *et al.*, 2010). However, the major challenge is to handle the bias caused by selective genotyping among the progeny tested candidate bulls, which reduces the ability of GEBVs to predict future breeding values.

### Data

Data from both conventional and genomic evaluations of 13 populations were submitted to Interbull Centre for the first official validation of the national genomic evaluations for protein yield. Some populations included more than one country and some populations were different breeds from the same country. Nine of the 13 populations were Holstein. Each national genetic evaluation centre was requested to submit 6 files:

- Conventional genetic data–Full dataset (***CF***)
- Phenotypic data–Full dataset (***DF***)
- Conventional genetic data–Reduced dataset (***CR***)
- Genomic data–Reduced dataset (***GR***)
- Description of GEBV test results (***F731***)
- Genotyping information on test bulls (***F732***)

where, full dataset corresponds to the current evaluation, and reduced dataset corresponds to the same bulls as in the full dataset, however based on evaluations from reduced phenotypic data, i.e., excluding the last four years of observations (four years to be consistent with Interbull validation method III (Boichard et al., 1995)).

*CF* file: containing current conventional national genetic evaluation information, as submitted to Interbull Centre for international genetic evaluation, including EBV, effective daughter contribution (EDC), and reliability ($r^2_{EBV}$).

*DF* file: containing current conventional national genetic evaluation, including either the de-regressed national EBV (DEBV) or DYD.

*CR* file: containing conventional genetic evaluation of the reduced dataset, including EBVr, EDCr and $r^2_{EBVr}$.

*GR* file: containing genomically enhanced breeding values (GEBV) (combined direct genomic and conventional genetic evaluations) of the reduced dataset, including GEBVr, GEDCr (GEDCr > EDCr), and $r^2_{GEBVr}$.

*F731* file: containing results of the validation test performed by the national genetic evaluation centre. The main purpose of this file was to check if the data editing has been done properly, and whether DEBV or DYD, GEBVr, and EBVr are on the same scale. In order to conduct the validation test these data must be expressed on the same scale.

*F732* file: containing genotyping information on test bulls (selection candidates for genotyping) representing the genotyped and the non-genotyped bulls. This file together with *CF* was used to estimate selection intensity for the genotyped candidates.

## Method

The method proposed by Mäntysaari *et al.* (2010) was used for the validation of national genomic evaluations. This method is similar to Interbull validation method III (Boichard *et al.*, 1995), which tests the consistency of national conventional genetic evaluations by successive evaluations. Therefore, successive genetic evaluations of the same bulls should have the same expectation close to their true breeding values. However, in the genomic validation method (Mäntysaari *et al.*, 2010), the slope of the regression of future EBV from the current GEBV is tested.

With successive evaluations, more information becomes available for the evaluation of the bulls. Although, the results of successive evaluations do not necessarily have to be equal to each other, genetic trends should remain stable by adding extra information from recorded daughters (Boichard *et al.*, 1995). Regressing estimated breeding values based on the current available information on the estimated breeding values from previously available information, the expected regression coefficients are 0 and 1 for the intercept and the slope, respectively (Boichard *et al.*, 1995). Consequently, the genetic trend of the same animals (i.e. the regression of estimated breeding values on the year of evaluation) should have a slope near to zero.

The method adopted for the validation of national genomic evaluations (Mäntysaari *et al.*, 2010) is as follow:

$$Y = b_0 + b_1 \times GEBVr + e$$

where, $Y$ is DEBV or DYD, $b_0$ is the intercept, $b_1$ is the linear regression slope, $GEBVr$ is the genomically enhanced EBV from the reduced dataset (EBVr), and $e$ is random residual effect.

This model shows how well the future breeding values can be predicted from GEBVr. The population of test bulls was assumed to be from a normal distribution of breeding values with selective genotyping based on higher early expectations for genetic merits (Figure 1). Selective genotyping causes bias in genomic evaluations. As a result of bias, the slope of the linear regression alters from the expected value of 1. If the non-genotyped animals are from the left tail of the distribution (Figure 1), $b_1$ as well as its expected value ($E(b_1)$) become less than 1, and if the non-genotyped animals are from the right tail of the distribution (e.g., for traits like calving interval, where a high EBV represents the inferiority of the animal), $b_1$ and $E(b_1)$ become greater than 1. Therefore, the aim was to test the hypothesis of $H_0$: $b_1 = E(b_1)$.

The model and the hypotheses were tested on the population of test bulls. This population had EDC greater than 20 (current evaluation) and EDCr equal to 0 (previous evaluation). This was because in the validation test, the available information before the bull gets its daughter performances are compared with the bull's evaluation including adequate information from daughter performances (Mäntysaari *et al.*, 2010). The population of test bulls was chosen from the bulls born since 2002, randomly sampled through an official AI scheme in the population of interest, and evaluated based on the first or both the first and the second crop of daughters in the population of interest.

When the dependent variable was DEBV, a weight statement was built in the regression model to minimize the weighted residual sum of squares $\Sigma_i w_i(Y_i - \hat{Y}_i)^2$ rather than the sum of squares.

where, $w_i$ is the weight, $Y_i$ is the observed $DEBV_i$, $\hat{Y}_i$ is the predicted $DEBV_i$ from the model, $w_i = EDC_i / (EDC_i + \lambda)$, and λ is the variance ratio.

Other options for *w* can be the reliability of $EBV_i$, or $EDC_i$ (current evaluation). The weighted regression model can be written in a matrix equation form as follows:

$$\begin{bmatrix} 1'w1 & 1'wX \\ X'w1 & X'wX \end{bmatrix}\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1'wY \\ X'wY \end{bmatrix},$$

where, *X* is the vector of GEBVr.

The three important results from the regression analyses were $b_0$, $b_1$, and $R^2$. The $R^2$ of the model using GEBVr was compared with the $R^2$ of a model in using EBVr in order to quantify the increase in the accuracy of prediction when genomic information are added to parent averages (Kistemaker and Sullivan, 2010; Mäntysaari *et al.*, 2010). The values of $b_0$ and $b_1$ were compared to 0 and $E(b_1)$, respectively.

Several steps were involved in the calculation of $E(b_1)$. Selection intensity (*i*), the proportion of the genotyped animals (*p*) and the deviation of the truncated data from the mean (*x*) can be estimated reciprocally (See Appendix Table A from Falconer and Mackay, 1996).

For the genotyped animals *i* was calculated according to Falconer and Mackay (1996):

$$i = S / \sigma$$

where, *S* is the selection differential and $\sigma$ is the total standard deviation based on the full dataset (*CF*), see Figure 1.

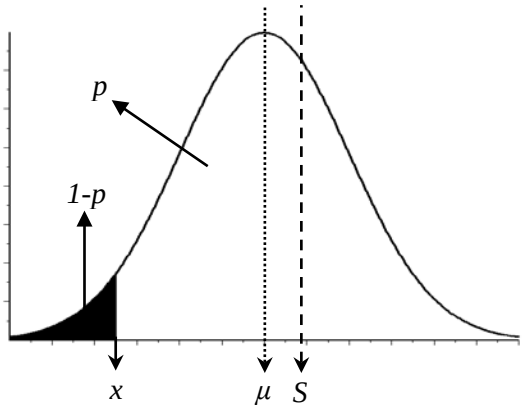


**Figure 1.** Diagram of the truncated normal distribution of breeding values, showing the proportion of the genotyped animals (*p*), selection differential (*S*), and deviation of the truncated data (*x*) from the mean ($\mu$).

With the knowledge of *i*, *p* can be read from Appendix Table A (Falconer and Mackay, 1996), or it can be calculated using a SAS program (see Appendix).

With either *i* or *p* known, the next step was finding the value of *x*. Again this value can be read from Appendix Table A (Falconer and Mackay, 1996) or from a known *p* value, using PERCENTILE function in Microsoft Excel (2007) or by the following command line in SAS (SAS Inst., 2010):

X = QUANTILE ('NORMAL', 1-þ)i

According to Mäntysaari et al. (2010):

$$E(b_1) = v_1 / v_2$$

where, $v_1$ is the genetic variance after selection, and $v_2$ is the genetic variance before selection.

Assuming $k = i(i - x)$; then

$$\frac{v_1}{v_2} = \frac{1-k}{1-kr^2}$$

where, $r^2$ is the accuracy before selection (the true accuracy given the estimated accuracy (accuracy after selection)), calculated as:

$$r^2 = \frac{R^2}{1-k(1-R^2)}$$

where, $R^2$ *is* the accuracy after selection, obtained from the regression analysis.

The null hypothesis $H_0$: $b_1 = E(b_1)$ was tested using a t-test, comparing the estimated *t* value against the critical *t* value:

$$t_{est.} = \frac{|b_1 - E(b_1)|}{SE(b_1)}$$

The critical *t* value can be found from a t-Table or using TINV function in Microsoft Excel, 2007 ($\alpha = 0.05$).

## Results and Discussion

From the 13 populations participated in the validation test, 11 of them could successfully pass the test (Table 1). In four populations, all the test bulls had been genotyped, therefore their *p* and $E(b_1)$ values were equal to 1. There were three other populations whose $b_1$ did not differ significantly from 1 in the presence of selective genotyping. All of these seven populations passed the test. Replacing the independent variable GEBVr by EBVr (Table 1, Model2), the observed and the expected slopes were significantly different for 8 populations. Even though no test has been recommended for EBVr, the results showed that parent averages alone are less informative than genomically enhanced parent averages to predict breeding values later in life, when daughter information is available. With the exception of two populations, the slopes for GEBVr were higher than for EBVr. This shows a closer scale of expression between young and proven bulls when using genomic information (Kistemaker and Sullivan, 2010).

For a population to pass the validation test of national genomic evaluation, despite of *i*, *p*, and *x* values (which are functioned in $E(b_1)$), $b_1$ should be as close as possible to $E(b_1)$. As the deviation between $b_1$ and $E(b_1)$ increases, larger $SE(b_1)$ values may allow populations to pass the test. However, larger $SE(b_1)$ values reduce the $R^2$ of the model and make the model less informative (i.e., overall, the national genomic evaluation is unbiased, but evaluations for some bulls may show a large deviation from the expected value). Therefore, a minimum threshold for $R^2$ is recommendable. However, because the size of the reference population with reduced data is considerably smaller than the actual reference population (Mäntysaari et al., 2010), very large $R^2$ values are not expected.

On average, using genomic information in addition to the pedigree information doubled the accuracy of evaluation (Table 1). The results also showed that even a weak selective genotyping (7.4% non-genotyped animals) can reduce $E(b_1)$ from 1 to 0.79 (Table 1, Population A). Therefore, national genomic evaluations might be considerably biased by ignoring information from non-genotyped animals.

It is important that different data, especially DEBV (or DYD), GEBVr and EBVr are on the same scale and follow a normal distribution, otherwise the assumptions of the method may become violated.

## Acknowledgment

The authors appreciate P. VanRaden (USDA, USA), Z. Liu (VIT, Germany) and E. Mäntysaari (MTT, Finland) for their constructive inputs in technical discussions.

## 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. 77,* 431–437.

Falconer, D.S. & Mackay, T.F.C. 1996. *Introduction to Quantitative Genetics*, Ed 4. Longmans Green, Harlow, Essex, UK.

Kistemaker, G.J. & Sullivan, P.G. 2010. Experiences in validating genomic evaluations. *Interbull Bulletin 40*, 235–239.

Mäntysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. Proceedings of the Interbull Meeting, March 4–5, 2010, Paris, France. *Interbull Bulletin 41*, 17-21.

Microsoft Office Excel. 2007. Microsoft©

SAS Institute Inc. 2010. SAS/STAT 9.2 *User's Guide*, Second Edition. Cary, NC: SAS Institute Inc.

Sullivan, P.G. & VanRaden, P.M. 2009. Development of genomic GMACE. *Interbull Bulletin 40*, 157–161.

**Table 1.** Summary of the results of the genomic validation test for protein yield, for different models and different participated populations

| Population | i | p | Model[a] | $b_1$ | $SE(b_1)$ | $E(b_1)$ | $R^2$ | t[b] |
|---|---|---|---|---|---|---|---|---|
| A | 0.152 | 0. 926 | 1 | 0.69 | 0.119 | 0.79 | 13 | **0.83** |
| | | | 2 | 0.31 | 0.095 | 1.00 | 4 | 7.30 |
| B | 0.013 | 0. 996 | 1 | 0.96 | 0.027 | 0.98 | 47 | **0.75** |
| | | | 2 | 0.89 | 0.040 | 1.00 | 27 | 2.86 |
| C | 0.079 | 0. 966 | 1 | 0.46 | 0.114 | 0.87 | 11 | **3.56** |
| | | | 2 | 0.30 | 0.123 | 1.00 | 4 | 5.66 |
| D | 0.000 | 1.000 | 1 | 1.03 | 0.022 | 1.00 | 54 | **1.44** |
| | | | 2 | 0.99 | 0.037 | 1.00 | 27 | 0.37 |
| E | 0.128 | 0.940 | 1 | 0.96 | 0.102 | 0.83 | 21 | **1.30** |
| | | | 2 | 0.96 | 0.144 | 1.00 | 12 | 0.26 |
| F | 0.037 | 0.986 | 1 | 0.76 | 0.062 | 0.94 | 23 | **2.83** |
| | | | 2 | 0.89 | 0.101 | 1.00 | 14 | 1.05 |
| G | 0.000 | 1.000 | 1 | 0.93 | 0.039 | 1.00 | 56 | **1.70** |
| | | | 2 | 0.92 | 0.049 | 1.00 | 44 | 1.67 |
| H | 0.110 | 0. 950 | 1 | 0.87 | 0.021 | 0.89 | 40 | **0.50** |
| | | | 2 | 0.76 | 0.029 | 1.00 | 21 | 8.38 |
| I | 0.102 | 0.954 | 1 | 0.89 | 0.031 | 0.90 | 44 | **0.20** |
| | | | 2 | 0.49 | 0.038 | 1.00 | 13 | 13.65 |
| J | 0.038 | 0.986 | 1 | 0.88 | 0.035 | 0.95 | 37 | **1.85** |
| | | | 2 | 0.86 | 0.054 | 1.00 | 19 | 2.64 |
| K | 0.000 | 1.000 | 1 | 0.96 | 0.038 | 1.00 | 49 | **1.07** |
| | | | 2 | 0.81 | 0.048 | 1.00 | 30 | 4.01 |
| L | 0.000 | 1.000 | 1 | 0.99 | 0.029 | 1.00 | 48 | **0.51** |
| | | | 2 | 1.00 | 0.054 | 1.00 | 21 | 0.06 |
| M | 0.108 | 0. 950 | 1 | 0.98 | 0.050 | 0.90 | 49 | **1.64** |
| | | | 2 | 0.88 | 0.053 | 1.00 | 40 | 2.35 |

[a]Independent variables for model1 and model2 are GEBVr and EBVr, respectively.
[b]Critical value = 1.96

**Appendix:** A SAS program for the estimation of *p* value

```
DATA est_p;
i = &iVALUE;
sign = 1;
IF i < 0 THEN sign = -1;
i = ABS(i);
density = 0;
sum = 0;
pi = CONSTANT ('PI');
e = CONSTANT ('E');
DO j = 5000 TO -5000 BY -1;
        density = density + 1/SQRT(2*pi) * e**(-0.5*(j/1000)**2);
        sum = sum + (j/1000) * 1/SQRT(2*pi) * e**(-0.5*(j/1000)**2);
        IF (sum / density > i) THEN DO;
            p = density / 1000;
        END;
END;
IF sign = -1 THEN p = 1 – p;
i = i*sign;
RUN;
```

Source: P. VanRaden (Personal communication), modifications by M.A. Nilforooshan
&iVALUE is known.