# An Approach to Account for Selection Bias in National Evaluations due to Genomic Selection

*C. Patry*[*†], V. Ducrocq[*]

## Introduction

Despite the rapid development of genomic selection in dairy cattle, there is a consensus on the need to maintain classical national and international evaluations. However, the inclusion of a pre-selection step based on genomic selection in breeding schemes invalidates some of the assumptions necessary to get optimal BLUP properties in evaluation systems. Indeed, considering only one generation of genomic selection, it has been shown that such a bias due to genomic pre-selection exists (Patry, C. and Ducrocq, V. (2009)): the estimated breeding values of pre-selected bulls and their daughters are underestimated and standard reliabilities are overestimated. Depending on the heritability of the trait and the applied selection intensity, the systematic difference between estimated and true breeding values may be substantial and bull ranking may be inaccurate. Thus, it appears necessary to correct for genomic pre-selection. This is particularly relevant at the international level when the use of genomic information in breeding schemes strongly varies across countries. A simple approach for the correction of a pre-selection bias is presented here. It consists in blending genomic information with classical performances into a joint BLUP analysis for all selection candidates. The potential bias was computed in various situations, considering different selection intensities, heritabilities and availability of pre-selection information.

## Material and methods

**Simulation of a genomic pre-selection.** Based on pedigree information and an assumed genomic equivalent daughter contribution (gEDC, Van Raden, PM., Van Tassell, CP., Wiggans, GR. *et al*. (2009)), true and genomically enhanced breeding values (TBV, GEBV) were simulated for a cohort of n sires (YS) of the French Holstein population (see Patry, C. and Ducrocq, V. (2009)). This cohort represents the youngest bulls for which daughters with conformation data were already available in the current evaluation files. Conformation was chosen as an example because of the simplicity of its genetic evaluation. gEDC represents the number of daughters that would lead to an increase in the sire's reliability equivalent to the one obtained via genomic evaluation.

In order to mimic a genomic pre-selection step with a selection intensity of 1/m, TBV and GEBV were generated m times for each YS, hence mimicking genomic evaluation of m full-sibs. Among these full-sibs, the highest GEBV was assigned to the YS undergoing progeny test whereas the remaining full-sibs were supposed to be culled. Performances for these

[*] INRA, UMR1313 - Génétique Animale et Biologie Intégrative, 78352 Jouy en Josas, France

[†] UNCEIA, 149, rue de Bercy, 75 595 Paris Cédex 12 France

sire's daughters were also simulated (see Patry, C. and Ducrocq, V. (2009) for details). The n selected YS – the ones with actual daughters in the evaluation file - were evaluated as in the national system based on their daughters' performance. This led to a situation where a bias is suspected [scenario with Genomic Pre-Selection = GPS]. In the reference scenario [REF], there was no pre-selection based on genomic information: the same n YS were evaluated but their TBV and GEBV were simulated only once (i.e., m=1).

**Computation of genomic equivalent daughter performances.** When genomic pre-selection existed, (n-1)m YS were culled without having daughters to be included in the evaluation. Consequently, the distribution of the estimated mendelian sampling terms of selected sires was no longer centered around 0, leading to biased BLUP solutions (Patry, C. and Ducrocq, V. (2009)). Two scenarios were studied where additional information from all genotyped candidates [COR scenario] or from the selected ones only [GPS+ scenario] was added to the national evaluation, to account for pre-selection. This additional information was expressed as "Genomic Equivalent Daughter Performance" ( $\mathrm{EDP}_i^G$ ) in such a way that, for scenario COR, the selection process was fully described in the data (Ducrocq, V. and Liu, Z. (2009)). For this step, genomic EDC ( $\Psi_i^G$ ) were supposed to be known as well as genomic values ( $\hat{a}_i^G$ ). $\mathrm{EDP}_i^G$ were calculated multiplying the coefficient matrix of the Mixed Model Equations (MME) by $\hat{a}_i^G$ :

$$\left[\begin{pmatrix}\Psi_1^G & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_m^G\end{pmatrix} + \alpha \mathbf{A}^{-1}\right]\begin{bmatrix}\vdots\\ \hat{a}_i^G \\ \vdots\end{bmatrix} = \begin{bmatrix}\vdots \\ \Psi_i^G\ \mathrm{EDP}_i^G \\ \vdots\end{bmatrix} \qquad (1)$$

**Breeding values estimation.** Four BLUP animal model evaluations were performed, including or not the genomic performances $\mathrm{EDP}_i^G$ of either all or some of the genotyped bulls (table 1). In the REF and GPS scenarios, all available field performances were included. In the other scenarios, $\mathrm{EDP}_i^G$ with their corresponding weight $\Psi_i^G$ were also added to the conventional BLUP mixed model equations, for all (COR scenario) or selected (GPS+ scenario) genotyped bulls. Note that because genomic EDC was expressed as daughter equivalents, they first needed to be transformed into equivalent own performances ( =(α/k) $\mathrm{EDP}_i^G$ where $\alpha$=(1-h²)/ $h^2$ and k=(4/$h^2$)-1 ) and genomic EDP were multiplied by 2 before being incorporated into the animal model.

**Table 1: Summary of the scenarios studied**

| Scenario | Genomic pre-selection | Inclusion of genomic performance | Inclusion of information on culled bulls |
|---|---|---|---|
| REF | no | no | no |
| GPS | yes | no | no |
| GPS+ | yes | yes | no |
| COR | yes | yes | yes |

Therefore, for the COR scenario, the EBV of the selected YS combined both classical and genomic performances whereas for culled YS, only genomic performances were used. In the GPS+ scenario, only selected bulls were evaluated using genomic performances: genomic EDP were computed for the selected YS only, disregarding information from culled bulls.

**Bias assessment.** For each YS, TBV and EBV were available, so that bias was assessed as the average difference between estimated and true breeding values, separately for selected young bulls and theirs daughters. Each scenario was replicated 20 times. The mean and standard deviation of the bias as well as reliabilities defined as the square correlation between EBV and TBV and the mean squared error of prediction (MSEP) were computed for the four scenarios: REF, GPS, GPS+ and COR.

**Numerical application.** The simulations were performed for two conformation traits for the Holstein breed: udder depth (UD) and foot angle (FA) with a heritability of 0.36 and 0.15 respectively. n=799 YS were identified as selection candidates for UD and n=601 YS for FA; 40,222 and 31,976 daughters were targeted. The full animal model evaluation included 5,917,701 animals and 4,110,229 records. Two selection intensities were considered: 25% (m=4) and 10% (m=10). gEDC were supposed to be equal to 10 for UD and 26 for FA for all genotyped sires. Given the heritability of the two traits, this corresponded to the same increase in reliability due to genomic information.

## Results and discussion

**Bias in breeding values' estimation.** Table 2 shows the average difference between EBV and TBV for udder depth, separately for selected young sires and their daughters: for both groups, BLUP solutions are significantly biased in the GPS scenario, as in Patry, C. and Ducrocq, V. (2009). This bias is virtually unchanged in the GPS+ scenario. When information from culled bulls is also included, the bias is then not significantly different from zero. Whatever the group of interest (young sires or their daughters), the selection intensity (Table 3) or the heritability of the trait, combining genomic with classical performances for all candidates (selected and culled ones), corrects for pre-selection bias.

**Table2: Mean, standard deviation of the bias, reliabilities and Mean square error of prediction (MSEP) for udder depth when 25% of the young sires are pre-selected based on their genomic evaluation (799 selected bulls; 20 replicates)**

| | **Young sires** | | | | **Daughters of the young sires** | | |
|---|---|---|---|---|---|---|---|
| | EBV-TBV[(a)] | $\rho^2_{(TBV,EBV)}$ | Rel[(b)] | MSEP | EBV-TBV[(a)] | $\rho^2_{(TBV,EBV)}$ | Rel[(b)] |
| **REF** | 0.005 ± 0.015 (ns) | 0.754 | 0.815 | 0.184 | 0.011 ± 0.006 (ns) | 0.413 | 0.476 |
| **GPS** | -0.153 ± 0.014 (***) | 0.713 | 0.815 | 0.186 | -0.047 ± 0.005 (***) | 0.386 | 0.476 |
| **GPS+** | -0.145 ± 0.014 (***) | 0.747 | 0.846 | 0.165 | -0.048 ± 0.005 (***) | 0.388 | 0.479 |
| **COR** | -0.015 ± 0.015 (ns) | 0.745 | 0.848 | 0.148 | -0.002 ± 0.006 (ns) | 0.391 | 0.479 |

(a): in genetic standard deviations ; (b) REL = approximate reliabilities computed from mixed model equations ; ***: $p<0.001$

**Quality of estimation.** In all cases, reliabilities of selected animals computed from mixed model equations –here, using Harris and Johnson's (1998) approximation - overestimates the EBV accuracy measured as $\rho^2_{(TBV,EBV)}$ in table 2. When young sires were pre-selected based on their genomic evaluation, this accuracy decreased and the mean squared error of prediction slightly increased. Combining genomic equivalent daughter performances with classical performances brought additional information which increased $\rho^2_{(TBV,EBV)}$, so the ranking among selected bulls was better. But ignoring information on culled bulls (GPS+) led to higher MSEP. Because of the bias in GPS and GPS+ scenarios, the overall ranking of young sires is not as good as in the COR scenario, for which EBV are no longer biased.

**Table 3: Mean and standard deviation of the bias in the cohort of young sires for udder depth (799 selected bulls, 10% of genomic pre-selection) and foot angle (601 selected bulls, 25% and 10% of genomic pre-selection) over 20 replicates**

| | **Udder depth** | **Foot angle** | |
|---|---|---|---|
| Selection intensity | 10% | 25% | 10% |
| **REF** | 0.005 ± 0.015 (ns) | 0.011 ± 0.018 (ns) | 0.011 ± 0.018 (ns) |
| **GPS** | -0.228 ± 0.017 (***) | -0.222 ± 0.019 (***) | -0.336 ± 0.024 (***) |
| **GPS+** | -0.219 ± 0.017 (***) | -0.220 ± 0.019 (***) | -0.332 ± 0.024 (***) |
| **COR** | -0.028 ± 0.016 (ns) | -0.001 ± 0.019 (ns) | -0.013 ± 0.023 (ns) |

## Conclusion

This simulation confirmed the existence of a bias when national evaluations do not account for pre-selection of young sires based on genomic information (Patry and Ducrocq, 2009). This bias is stronger among sires than for their daughters. It has an impact on rankings, on true and computed reliabilities and on MSEP. A method converting genomic EBV of genotyped animals into genomic performances ( $\mathrm{EDP}_i^G$ ) which can then be included in national evaluations was shown to correct the bias under certain assumptions. Two important assumptions are: a) the EBV and GEBV are for "the same trait", with the same genetic variance. Note that if GEBV explain only a fraction of the total genetic variance, $\mathrm{EDP}_i^G$ can still be combined with real performances but as a correlated trait. b) information of all genotyped animals must be included. This supposes that GEBV of culled bulls must be available at national and international evaluation centers. The simplicity of the approach also rests on the fact that there is no need to modify existing evaluation software.

## References

Ducrocq, V., Liu, Z. (2009). *Interbull Bulletin*, 40:172–177.

Harris, B., Johnson, D. (1998). *J. Dairy Sci.,*81:2723-2728.

Patry, C., Ducrocq, V. (2009). *Interbull Bulletin*, 40:167–171.

Van Raden, PM., Van Tassell, CP., Wiggans, GR. *et al*. (2009). *J Dairy Sci.*, 92:16-24.