Implementation of Genomic Selection at National Level: Impact of Pre-Selection and Biased National BLUP Evaluations on International Genetic Evaluations

Clotilde Patry [1,2], Hossein Jorjani[3] and Vincent Ducrocq[1]
[1]INRA UMR 1313, Génétique Animale et Biologie Intégrative, 78 352 Jouy-en-Josas, France
[2]Union nationale des Coopératives d'Elevage et d'Insémination Animale (UNCEIA), 149, rue de Bercy, 75 595 Paris Cédex 12 France
[3] Interbull Centre, Department of Animal Breeding and Genetics SLU, Box 7023, S-75007 Uppsala, Sweden
e-mail: clotilde.patry@jouy.inra.fr

Abstract

It has been shown that genomic pre-selection of young bulls leads to biased national BLUP evaluations if information of culled animals is not accounted for. The objective of this study was to assess the consequences of such missing and/or biased national data on international genetic evaluations. Various genomic selection scenarios were simulated in 3 actual populations participating in Interbull evaluations. They were first simulated separately to clearly understand how bias is propagated. Then, they were combined to illustrate a more realistic case. The current international genetic evaluations when national proofs are available for all candidates and supposed to be unbiased were used as a reference. Hence, bias was measured among young sires and by country of origin as the average difference between the current evaluations and the ones obtained in the simulated scenarios. Bias due to missing information on the culling process at national level or due to national biased proofs in one country highly penalized young sires from that country. But it also had an impact on the evaluation of foreign young and older sires. Moreover, it becomes more difficult if not impossible to predict this impact when different sources of bias were combined from different countries: all effects interact with each other. But a change in ranking is certain so that selection efficiency becomes clearly suboptimal.

Keywords: international genetic evaluations –genomic selection - selection bias

1. Introduction

Since 2008, genomic selection has been implemented in several countries participating in Interbull evaluations. Various genomic selection strategies are being adopted across countries but also within country. In the male population, molecular information is mainly used to implement a pre-selection step of the young sires to be either combined with progeny testing or to allow an immediate use of young sires. It follows that young sires which are eliminated after this pre-selection step have no daughters with performance and the expected Mendelian sampling contribution is no longer zero among the selected ones. Consequently, two assumptions of the mixed model methodology, on the completeness of information about selection decisions and on Mendelian sampling distribution are violated, at national and also at international level, when MACE methodology (Schaeffer, 1994) is used.

Patry and Ducrocq (Patry and Ducrocq, 2011a) showed that under such circumstances the national BLUP evaluations were biased due to genomic pre-selection. Breeding values of the selected young sires and their daughters tended to be underestimated. It is thus essential to account for this pre-selection step in national genetic BLUP evaluations. Two alternative solutions have been proposed, based on a single step approach (Aguilar *et al.*, 2010; Christensen and Lund, 2010) or the inclusion of genomic pseudo-performances into BLUP evaluations as described by Ducrocq and Liu (2009) and implemented by Patry and Ducrocq (2011b).

Such issues might also have an impact at the international level. Only a part of the

participating countries in Interbull evaluations have implemented genomic selection. Among them, some may account for genomic pre-selection at national level and others not. Furthermore, national BLUP evaluations are currently sent to Interbull for sires with daughters, i.e., for selected young sires only. Without any changes of national and international evaluation practices, it is feared that international genetic evaluation might be biased. The objective of this study was to assess the possible consequences on international genetic evaluations due to non random missing data in MACE and due to bias transmission through the international genetic relationships and genetic correlations between countries.

2. Materials and Method

General strategy – Simulations were based only on estimated breeding values (EBV) available at Interbull and do not include any genomic information. For young bulls (the youngest cohort of bulls with EBV available), it was considered that these EBV were equivalent to genomically enhanced breeding values (GEBV) or direct genomic values (DGV) that may be or may be not sent at Interbull for some or all of them. In other words, national genomic pre-selection was mimicked by assuming that some of the EBV of the youngest cohort of sires were not sent to Interbull.

The characteristics of these young sires' EBV and their availability at Interbull level were simulated (1) by excluding some members of each half-sib family to mimic genomic pre-selection and (2) by biasing national proofs of the selected candidates. The consequences of the implementation of genomic pre-selection on international breeding values were assessed as a bias comparing MACE results on domestic and foreign scales with the ones obtained from the original file. The latter were supposed to be unbiased, corresponding to the control (CTL) scenario.

Dataset – Data were national genetic evaluations for protein yield in Holstein breed. All the data required for the August 2010 Interbull routine evaluation were available. However, we only used data from 3 countries hereafter called A, B, and C. These included 57,688 sires out of about 120,000 bulls from the 27 countries participating to this international evaluation. Heritabilities were 0.48 in country A, and 0.30 in countries B and C. All EBV were scaled to their domestic genetic standard deviation. Genetic correlations were 0.85 between A and B, 0.87 between A and C and 0.90 between B and C.

Cohort of interest **-** Sires born between 2003 and 2006 and having only daughters in their country of origin defined the young sire cohort "YS". In each country, we focused on sire families including at least 10 half-sibs. There were 2,234 such young sires in country A, 1,310 in B and 3,602 in C. They had on average 107 recorded daughters for county A, 69 for country B and 97 for country C.

Simulations **-** Three types of scenarios were simulated. First, in the "SEL" scenarios, 10% of the YS were retained. Pre-selection was implemented in only one country, either A, B or C, or in the three countries at the same time. These scenarios were called SEL-A, SEL-B, SEL-C and SEL-all, respectively. For simulations, the selection criterion was actually based on the Mendelian sampling estimates within family and country. These estimates were computed as the difference between EBV and parent average (PA) from the CTL international evaluations. Impacts on rankings were also studied. Change in proportion of YS from each country among the top 100 sires was compared between the simulated scenarios and the CTL situation.

A second type of scenario was considered where genomic pre-selection of YS was not accounted for at national level in one country, implying that this country (A, B or C) sends biased national proofs to Interbull (BNP-A, BNP-B, BNP-C scenarios – BNP for "Biased National Proofs"). Here it was assumed that data for all candidates were sent to Interbull, i.e., no selection was implemented; information on culled animals is available at Interbull. The bias (Δ_i) was drawn as a random standard normal variable for each young sire (i) from a normal distribution *N(-0.227, 0.016)* and added to each actual national breeding value (y_i). These values of mean and variance were chosen from the study of Patry and Ducrocq (2011a). The sum $y'_i = y_i + \Delta_i$ was then considered as the new input for international genetic evaluations.

Third, in the "CMB" scenarios, the effects of partial (selected) transmission of data and biased national proofs on international evaluations were considered together. The three countries were assumed to have implemented genomic selection and to all send data on selected young sires but only one did not account for genomic pre-selection at national level and therefore for that country, the national proofs received by Interbull were biased. This later country was either A, B or C leading to CMB-A, CMB-B or CMB-C scenario. Scenarios BNP and CMB were replicated 10 times.

3. Results

To illustrate the major effects of genomic pre-selection use on international genetic evaluations, only results on 4 scenarios (SEL-A, BNP-A, SEL-all, and CMB-A) are presented here but conclusions were drawn from all simulation results.

Effect of pre-selected data from one country – When country A was assumed to send only a reduced set of national proofs to Interbull after genomic pre-selection of young sires (SEL-A scenario), the EBV of young sires coming from A tended to be penalized compared with foreign YS. Because of pre-selection, the average EBV across domestic YS was higher and their standard deviation was smaller than those observed among foreign YS. Depending on which scale the international evaluation was expressed, this contrast made the EBV of domestic YS underestimated or the EBV of foreign YS overestimated: on their local scale, international proofs of domestic YS are barely changed - most of the information was local - so that bias was virtually null and EBV of foreign YS was overestimated. However, EBV of domestic YS were clearly underestimated on foreign scale (see Table 1 and Graph 2).

Table 1. Mean bias (in genetic standard deviation) among young sires by country of origin when country A sends partial (pre-selected) data to Interbull (Scenario SEL-A).

Country of origin	A scale	B scale	C scale
A	-0.01	-0.11	-0.10
B	0.15	0.01	0.03
C	0.17	0.02	0.00

Effect of using biased domestic proofs from one country in MACE - Country A was assumed to send to Interbull a complete list of national proofs but these included biased proofs for young sires (BNP-A scenario). YS from country A were as underestimated on A scale as they were at national level. On B and C scales, they were also underestimated, but to a lesser extent. YS from countries B and C were also underestimated on the A scale (see Table 2 and Graph 3). This was expected: on domestic scale, the most important contribution to international EBV is the de-regressed national EBV, i.e., the daughters' average performance. Here, de-regressed national EBV were actually biased so that resulting international EBV were also biased. On foreign scales, for domestic YS, the contribution of daughters' performances, i.e., the de-regressed biased proofs were less important - it was "regressed" according to the genetic correlation between countries - and the contribution of parent average (PA) increased so that bias on foreign scale was buffered for domestic YS. The magnitude of bias on foreign C scale was higher than on foreign B scale, probably because genetic correlation between A and C was higher than between A and B. On each scale, bias among YS from A is transmitted to YS from B and C through the genetic relationship matrix.

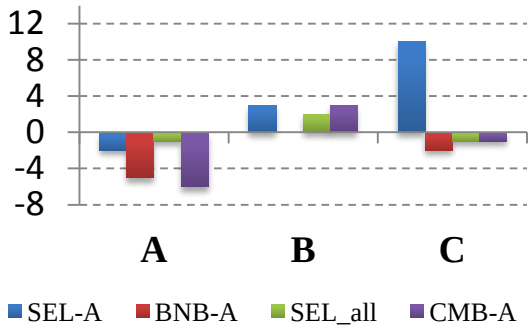
Table 2. Mean bias (in genetic standard deviation) among young sires by country of origin when country A sends biased data (Scenario BNP-A) to Interbull.

Country of origin	A scale	B scale	C scale
A	-0.22	-0.13	-0.14
B	-0.11	0.00	-0.03
C	-0.11	-0.02	0.00

Impact on rankings **–** Changes in proportion of YS from each country among the top 100 sires between the simulated scenarios and the CTL situation were depicted by Graph 1. When country A was assumed to send pre-selected data to Interbull (SEL-A scenario), the number of YS in the top 100 sires on the A scale increased from 54 to 65. Actually, some domestic YS and older sires were excluded and replaced by foreign YS as these were overestimated. On foreign scales, the proportion of YS remained stable except that few YS from A were replaced by foreign YS. When A was assumed to send biased domestic proofs to Interbull (BNP-A scenario), the proportion of YS decreased in favour of older sires. On the scale of the country sending biased national proofs (A in this case), all YS, irrespective of their country of origin, were penalized. But country A in which bias was the highest was particularly penalized with 5 YS fewer (out of 18 initially) in the top 100. On foreign scales, there were almost no changes. Only few YS from A were removed from the top 100 sires.

Implications of genomic selection **–** Now consider a more realistic scenario where all 3 countries implemented genomic pre-selection and accounted for it at national level but sent only partial information (on selected YS only) to Interbull (SEL-all scenario). Furthermore, consider that one country (A for example) did not account for genomic selection at national level (CMB-A scenario). In such a case, the bias observed in scenario BNP-A tended to be added to the biases observed in scenario SEL-all. Finally, YS from country A were the most penalized. However, YS and rankings were affected irrespective of the country of origin.

Graph 1. Change in number of YS from each country among the top 100 sires for SEL-A, BNB-A, SEL-all and CMB-A scenarios.



Global trends from all simulated scenarios:

1) YS from the country sending incomplete data were penalized on all scales. Moreover, on the local scale, pre-selection clearly favored foreign YS and hindered older sires.

2) Even if only one country send biased national proofs, all YS were penalized, whatever the country of origin. However, YS from the country sending biased national proofs were the most affected, on all scales. Older sires were thus favored as well as foreign YS on foreign scales. But it also had an impact on foreign sires and caused a lot of re-ranking.

3) Bias due to sending partial data to Interbull or due to national biased proofs highly penalized YS from the country/countries responsible for these practices. It is more difficult to predict the impact on YS when different sources of bias were combined. All effects interact with each other. Moreover, genomic selection intensity might be different from one country to another, involving different proportions of missing data and magnitude of national bias. But change in rankings is certain.

4. Discussion

This study aimed at assessing the consequences of new national practices such as genomic pre-selection of young sires. Strategies might vary a

lot between countries but also within country. Consequently, it is difficult to predict the magnitude of bias. With more countries implementing genomic selection but also with more generations of genomic selection, the situation will be come very messy and complex. However, it clearly appeared that missing and biased data at Interbull level leads to biased estimated breeding values in such a way that rankings, market shares and selection decision would be affected.

All countries participating in Interbull evaluations are connected through the MACE evaluations. Even if some countries are not implementing genomic selection or, at the other extreme, are using it but are already accounting for it in their national evaluations, they still might see their tools for selection decision affected and this issue can definitely not be ignored. One solution is that all countries implementing genomic selection should first account for it at each national level and then send all available and unbiased data to Interbull. Methodologies have been already proposed (Aguilar *et al.,* 2010; Patry and Ducrocq, 2011b) However, several complications may exist.

First, only GEBV are available for culled candidates so that GEBV and EBV should somewhere be combined in MACE evaluations. It is well recognized that care should be taken to avoid residual correlations when 2 animals get GEBV in 2 different countries. This is being considered in the GMACE procedure (VanRaden and Sullivan, 2010; Zumbach *et al.,* 2011).

Second, international genetic evaluations are particularly used in genomic prediction equations when reference populations include animals from several countries. It is then feared that genomic information will be double counted.

Third, including all available information involves having access to all genotyped animals from all countries participating in Interbull evaluations, which is a formidable political challenge in itself. The technical issue is to manage such a massive quantity of data which is very likely to increase very quickly.

Fourth, we wondered about the meaning of the current validation tests of national genetic evaluations. They are based on yearly genetic trend (Interbull test I, II, III) (Boichard *et al.,* 1995) or on Mendelian sampling estimate (Interbull test IV) (Fikse *et al.,* 2005). Both types are clearly impacted by genomic selection practices. It will become ever more difficult to respect validation rules and in addition to detect other types of bias as the ones described in this study.

There is an urgent need to find (inter)national genetic evaluations and validation procedures much more robust to the vast heterogeneity of situations that will exist in the near future.

5. Conclusion

Genomic selection induces a clear cut-off point in genetic trends. In return, genetic evaluation systems must be adapted. It is very important that each country adapts its national evaluation system and accounts for genomic pre-selection. It follows that the Interbull community must also rework validation procedures of national evaluations. Delivering useful international genetic evaluations may not be threatened only if each participating country follows basic rules which are first useful at national level.

6. References

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. *J. Dairy Sci. 93,* 743-752.

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.

Christensen, O. & Lund, M. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol. 42,* 2.

Ducrocq, V. & Liu, Z. 2009. Combining genomic and classical information in national BLUP evaluations. Proceedings of the 2009 Interbull meeting. Barcelona, Spain. *Interbull Bulletin 40,* 172-177.

Fikse, F., Liu, Z. & Sullivan, P. 2005. Tolerance value for validation of trends in genetic variances over time. Proceedings of the 2005 Interbull meeting, Uppsala, Sweden. *Interbull Bulletin 33,* 200-203.

Patry, C. & Ducrocq, V. 2011a. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci. 94,* 1011-1020.

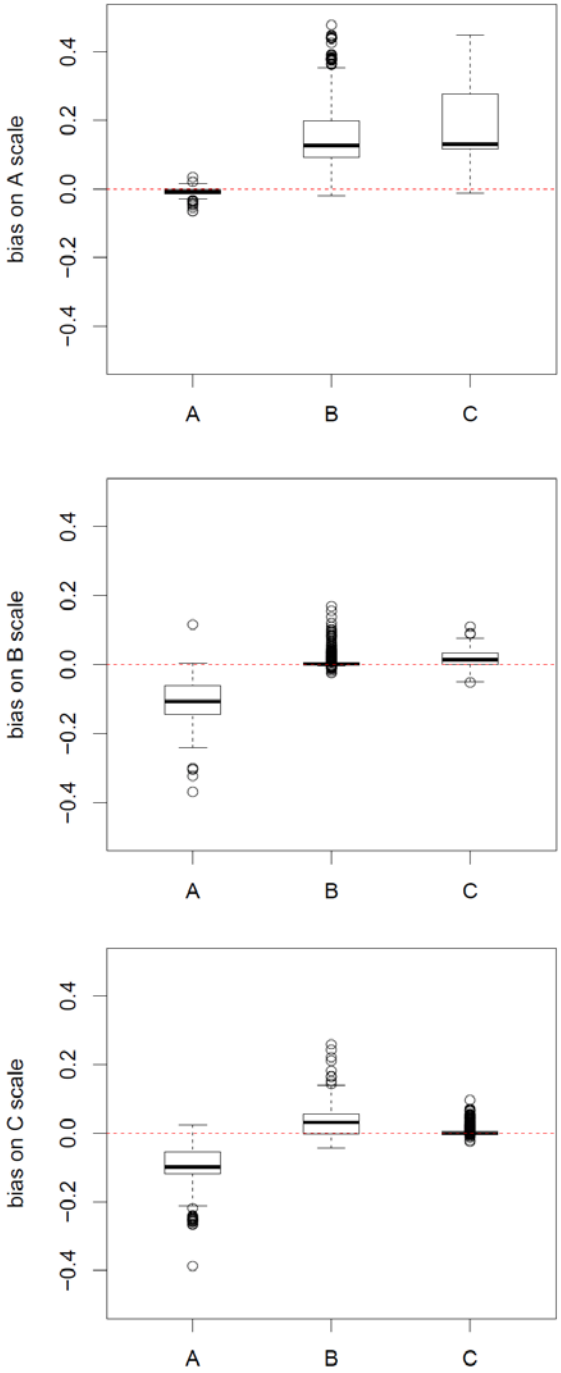
Patry, C. & Ducrocq, V. 2011b. Accounting for genomic pre-selection in national BLUP evaluations in dairy cattle. *Genet. Sel. Evol. 43,* 30.

Schaeffer, L. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci. 77,* 2671-2678.

VanRaden, P. & Sullivan, P. 2010. International genomic evaluation methods for dairy cattle. *Genet. Sel. Evol. 42,* 7.

Zumbach, B., Jakobsen, J., Forabosco, F., Jorjani, H. & Dürr, J. 2011. Data Selection and Pilot Run on Simplified Genomic MACE (S-GMACE). Interbull technical workshop Establishing the framework for international genomic evaluations, Guelph, Canada. *Interbull Bulletin 43,* 11-18.

Graph 2: Distribution of bias among young sires on the 3 scales and by country of origin when country A send pre-selected data (Scenario SEL-A)



Graph 3: Distribution of bias among young sires on the 3 scales and by country of origin when country A send biased data (Scenarion BNP-A)