MTEDC Software Available for Standardized EDC Calculations

P.G. Sullivan
Canadian Dairy Network, Canada

Abstract

Calculation of weighting factors for MACE analyses of complex traits requires extensive theoretical and programming developments, which are difficult and costly to duplicate by all countries participating in MACE. As a result, the evolution and implementation of advanced methods for calculating weighting factors has lagged behind recent developments in national genetic evaluation systems. Generalized EDC software was therefore created for use by all countries to facilitate a more efficient evolution of applied EDC methods in the future. Options currently available with the software and future planned developments are described. Applications are presented for two large data sets and considering two alternative EDC methodologies. Initial release of the software is planned for fall 2007.

1. Introduction

Effective daughter contributions (EDC), the weighting factors used in MACE, are approximated separately by each country contributing data to Interbull. Differences among methods or among the relative suitability of methods used to approximate EDC can adversely affect international comparisons in MACE, particularly for newly proven young sires. Approximation errors in EDC affect estimates of country sire variances and the subsequent scaling of Mendelian Sampling estimates among the evaluation scales of different countries. Foreign-scale MACE conversions for young sires originating from the same country as their parents are most notably affected (Fikse *et al.*, 2001).

Ideally, the EDC methods used by each country should be as accurate as possible. However, in practice, differences in EDC accuracy among countries can cause problems in MACE, and there will be greater differences in EDC accuracy if standardized and appropriate methodology is not consistently applied by all countries. Therefore, to optimize MACE results it is imperative to not only use the best EDC methods available, but to also ensure that methods are consistently applied by all participating countries.

Independent customized programming of EDC methods by each country increases the difficulty of standardizing and improving methods over time, because any changes require a duplication of programming efforts among dozens of genetic evaluation centers around the world. New EDC methods are required for some traits evaluated by Interbull, but suitable methodology improvements have required extensive and complicating programming efforts (Sullivan *et al.*, 2006), and would be difficult and costly for all countries to duplicate. Therefore, the purpose of the present paper was to initiate a system that can facilitate continuous improvement of EDC methods, while minimizing duplication of programming efforts and maximizing the consistency of application by each country participating in MACE.

2. Materials and methods

The EDC methods presented by Sullivan *et al.* (2006) were more accurate than the Interbull (2000) methods currently being used to calculate EDC for existing MACE applications, and also allow for calculation of EDCs under models too complicated for the current Interbull methods. Further improvements to the newer methods are also envisaged. The current version of EDC software allows for the application of either the current or the newer EDC methodology, without requiring any changes to input data files.

Applications of the software for MACE have so far been limited to a series of traits in Canada and Italy, selecting the current EDC methods (E2000) as required by Interbull.

Future updates to the EDC methodology applied in Canada and Italy will require minimal (almost zero) programming efforts by these two countries. User options available in the current version of EDC software include:

- single-trait or multiple-trait models for any number of traits.
- all or a subset of traits with direct and maternal genetic effects.
- direct and/or maternal permanent environmental effects.
- fixed or random contemporary group effects.
- heterogenous residual variances
- sire/dam or sire/mgs pedigree.
- EDC for any number of linear combinations of genetic effects.

Some improvements have been recently implemented, and final testing is nearing completion, for models involving multiple traits, maternal effects and permanent environmental effects, and random contemporary group effects.

The current version of software and future updates will be made generally available from CDN via ftp over the internet. A gnu or equivalent C compiler (comes standard with linux) will be needed to create a binary executable that will run on your system.

To apply the EDC software requires the following input files:

1. A parameter file specifying: the number of traits and linear functions of interest; the covariances among traits for all random effects in the model, which may include direct and maternal genetic and permanent environmental effects, contemporary group effects, and residual effects; the type of model (animal or sire); the names of all other input and output files.
2. A pedigree file.
3. A performance data file.

Animals must be numbered sequentially and consistently between the pedigree and performance files, but do not have to be numbered chronologically. Contemporary groups must be numbered sequentially. There is no sort requirement for the performance data file.

3. Results and Discussion

3.1. Run time and RAM requirements

Preliminary results for selected applications are shown in Table 1. Run times have been quite reasonable and are of minor concern relative to much longer run times generally needed for variance component estimation or prediction of breeding values. As was expected, the improved EDC methods (E2006) required more run time and computer memory (RAM) than the current Interbull methods (E2000), but were clearly feasible for the one large-scale example studied so far.

Table 1. EDC software performance.

Animals (Mil)	3.2	5.4	
Records (Mil)	2.1	3.7	
Traits[a]	20	5	
CPU (Ghz)	3.6	3.0	
EDC method	E2000	E2000	E2006
RAM (Gbytes)	7.8	3.0	3.5
Runtime (min)	80	4	8

[a]Each trait had 1 genetic effect (animal).

3.2. Comparison of methods

When applied to the 5-trait survival model in Canada, differences in EDC based on E2000 versus E2006 (Table 2) were consistent with the simulation results of Sullivan *et al.* (2006), and the explanations by Liu *et al.* (2001) of upward bias in EDC based on E2000. Contributions from correlated traits to EDC were reduced considerably with E2006. Among individual survival traits, the upward bias for E2000 was greater for early survival, due to lower culling rates and heritabilities and subsequently greater influence of correlated trait information on the earlier survival traits. Heritabilities were .007, .015, .022, .041, .054 for survival traits 1 through 5, and .098 for the linear function (LF) of the 5 traits.

The overestimation by E2000 relative to E2006 for EDC of individual survival traits was not observed for LF. E2000 gave smaller

EDC than E2006 for LF. Consequently, the LF results for E2000 were inconsistent with the individual trait results. For E2006, however, the LF results were consistently intermediate to the individual trait results, where expected.

The results in Table 2 were heavily influenced by the small percentage of sires that had thousands of daughters. In terms of MACE results, EDC are much less important for these sires than for young sires. Results were therefore repeated for (93% of) sires with a maximum 100 progeny records (Table 3). For these sires, E2006 were slightly higher than E2000 rather than lower, and the maximum EDC was much higher for E2006. These differences were due to the fact that E2006 accounts for grand-progeny records through daughters, which are ignored by E2000. Both E2006 and E2000 ignore grand-progeny records through sons to avoid double counting of records in MACE, but the same argument does not apply to daughters. Accounting for grand-progeny records through daughters is a critical advantage of E2006 over E2000 for models with maternal effects on traits such as calving ease.

It is important to note that differences between E2000 and E2006 depend very much on the choice of trait, or the linear function used to define the trait submitted to Interbull for MACE. Differences also depend on the specific application (genetic model and data structure) in each country.

3.3. Future developments

Additional features being developed for the EDC software include:

- reliability approximations for all animals in the pedigree.
- low-RAM option for huge applications.
- reduced animal models that fit only parent effects on records.
- models that include random genetic regression effects.
- improved adjustments for the estimation of contemporary group effects when contemporaries are genetically related.

4. Conclusions

The majority of software development needed for the initial release of a general-purpose EDC program that can be easily applied by all countries participating in MACE has been completed. Applications testing should continue but is also nearing completion. A software release is possible as early as fall 2007. Once countries have incorporated the software into their genetic evaluation systems, it will be easy for every country to apply both E2000 and E2006 for a more detailed application research study, prior to routine implementation of E2006 in MACE. Results from E2006 can replace progeny counts for calving ease and E2000 results for all other traits that do not involve random regressions of animal effects. A future upgrade of the EDC software can be considered for random regression models.

Acknowledgements

The pioneering efforts of Stefano Biffani (ANAFI) and Asheber Sewalem (CDN/AAFC) to incorporate the EDC software into the national genetic evaluation systems in their respective countries are appreciated.

References

Fikse, W.F. & Banos, G. 2001. Weighting factors of sire daughter information in international genetic evaluations. *J. Dairy Sci. 84*, 1759–1767.

Interbull, 2000. New weighting factors for the international genetic evaluation; revised July. Mimeo.

Liu, Z., Reinhardt, F. & Reents, R. 2001. The effective daughter contribution concept applied to multiple trait models for approximating reliability of estimated breeding values. *Interbull Bulletin 27*, 41-47.

Sullivan, P.G., Liu, Z. Jakobsen, J.H. & Fikse, W.F. 2006. More on weighting factors for complicated models. *Interbull Bulletin 35*, 112-116.

Table 2. Distributions of progeny counts, for survival in Canada, and EDC for sires with 1 or more progeny observations, by trait.

		Survival trait					LF[a]	
Statistic	**Variable**	**1**	**2**	**3**	**4**	**5**	**any**	**all**
N	Sires	50951	49731	46534	42218	35558	50951	35558
Mean	#Progeny	73	71	65	56	42	73	42
	E2000	511	226	167	94	77	38	54
	E2006	297	148	115	70	51	47	65
Standard deviation	#Progeny	820	794	726	606	431	820	431
	E2000	5942	2592	1852	1010	769	444	530
	E2006	1394	977	790	564	375	537	641
Maximum (*1000)	#Progeny	59	56	50	41	26	59	26
	E2000	417	179	124	65	45	31	31
	E2006	64	54	43	32	20	35	35
Correlation	#Progeny,E2000	.984	.991	.998	.998	.991	.987	.987
	#Progeny,E2006	.912	.975	.980	.990	.990	.994	.994
	E2000,E2006	.912	.976	.982	.991	.981	.992	.993

[a]Relative weights for the 5 traits in linear function LF were 29:24:20:17:10. Sires included had progeny observations for any trait (1st column) or all 5 traits (2nd column).

Table 3. Distributions of progeny counts, for survival in Canada, and EDC for sires with a maximum 100 progeny and 1 or more progeny observations, by trait.

		Survival trait					LF[a]	
Statistic	**Variable**	**1**	**2**	**3**	**4**	**5**	**any**	**all**
N	Sires	47606	46386	43206	39063	32669	47606	32699
Mean	#Progeny	16	15	14	12	9	16	9
	E2000	104	46	34	19	16	6	11
	E2006	123	53	39	22	17	8	13
Standard deviation	#Progeny	24	23	20	17	12	24	12
	E2000	180	78	55	30	22	13	15
	E2006	214	88	65	34	23	19	21
Maximum	#Progeny	100	100	100	83	73	100	73
	E2000	968	413	286	153	111	73	73
	E2006	4440	1875	1399	720	454	574	574
Correlation	#Progeny,E2000	.943	.963	.975	.973	.954	.948	.948
	#Progeny,E2006	.867	.891	.898	.893	.814	.838	.825
	E2000,E2006	.926	.927	.920	.918	.823	.877	.865

[a]Relative weights for the 5 traits in linear function LF were 29:24:20:17:10. Sires included had progeny observations for any trait (1st column) or all 5 traits (2nd column).