ELSEVIER

Reproducibility of the HERITAGE Family Study Intervention Protocol: Drift over Time

E. WARWICK DAW, PhD, MICHAEL A. PROVINCE, PhD, JACQUES GAGNON, PhD, JEAN-PIERRE DESPRES, PhD, CLAUDE BOUCHARD, PhD, ARTHUR S. LEON, MD, JAMES S. SKINNER, PhD, JACK H. WILMORE, PhD, AND D. C. RAO, PhD

PURPOSE: The primary goal of the HERITAGE Family Study was to document the role of the genotype in the response to aerobic exercise training. Toward this end, nuclear families were enrolled in a 20-week exercise training program, with a large variety of tests performed before and after the training. Since study drift has the potential to adversely affect the results, reproducibility and potential bias over six consecutive 4-month periods were examined for selected test.

METHODS: Intraclass correlations (ICC), technical errors (TE), coefficients of variation within subject (CV), and means were calculated with use of the pretraining test results for each of the six time periods. To check for homogeneity, hypothesis tests were performed on the intraclass correlations and means. If homogeneity was not found across all six periods, further tests were performed to assess differences between pairs of time periods.

RESULTS: There was little evidence for real drifts in reproducibility, with most tests having ICCs of 0.8 or better. Only a few tests showed any change over time, and in no case was there evidence of a systematic drift in mean values.

CONCLUSIONS: Overall, the reproducibility of the HERITAGE Family Study tests and assays considered in this paper was found to be very good, with no evidence of any systematic drift over time. *Ann Epidemiol 1997;7:452–462.* © 1997 Elsevier Science Inc.

KEY WORDS: Reproducibility, Bias, Quality Control, Exercise Test, Blood Lipids, Anthropometry, Blood Pressure, Multicenter Studies, Longitudinal Studies.

INTRODUCTION

The HERITAGE Family Study is a multicenter clinical exercise trial involving families. The objective of this study was to assess the response to a 20-week exercise program with particular interest in assessing the role of genetic factors in the response (1). The accuracy of that assessment will largely depend on the reproducibility of the measurements. While all measurements are made with some degree of error, appropriate quality assurance/quality control procedures can be used to quantify, explain, and minimize the magnitude of error and thus its effect on results.

Traditionally, two types of error are distinguished in quantitative measures: bias and precision. Bias, or systematic error, is the degree to which there is a systematic deviation from the "true" underlying value. For example, an improperly calibrated scale might produce values that are, on average, 10 g too high. A standard measure of bias is obtained by comparing the mean of a series of measurements to the mean of the same series of measurements under a "gold standard." Where a "gold standard" is lacking, sometimes the means of measurements under a series of different protocols are compared. Measurement error, also called precision or reliability, is the degree to which a measurement can be replicated. Precision might be quantified by the average standard deviation of a series of repeated measures. If they are poor, both bias and precision can have an adverse effect on the results of a study.

Adding to the complexity of quality assurance/quality control is the potential change in error over time, or study drift. Study drift may occur for a variety of reasons (2). Increasing error over time might result from decreasing attention to detail, new personnel receiving inadequate or different training, a measurement device becoming less accurate as it aged, or staff simply falling into a more familiar, casual pattern of doing things other than that specified in the standardized protocol. Decreasing error over time may also be observed, most frequently as a result of a "training"

From the Division of Biostatistics, Washington University Medical School, St. Louis, MO (E.W.D., M.A.P., D.C.R.); Physical Activity Sciences Laboratory, Laval University, Quebec, Canada (J.G., J-P.D., C.B.); School of Kinesiology and Leisure Studies, University of Minnesota, Minneapolis, MN (A.S.L.); Department of Kinesiology, Indiana University, Bloomington, IN (J.S.S.); Department of Kinesiology and Health Education, University of Texas at Austin, Austin, TX (J.H.W.); and Departments of Psychiatry and Genetics, Washington University Medical School, St. Louis, MO (D.C.R.).

Address reprint requests to: Dr. D.C. Rao, Washington University School of Medicine, Division of Biostatistics, Box 8067, 660 South Euclid Avenue, St. Louis, MO 63110-1093.

Received November 25, 1996; accepted May 16, 1997.

1047-2797/97/$17.00
PII S1047-2797(97)00082-3

Selected Abbreviations and Acronyms

LDL-cholesterol = low-density lipoprotein cholesterol
VLDL-cholesterol = very-low-density lipoprotein cholesterol
HDL-cholesterol = high-density lipoprotein cholesterol
TE = technical error
ICC = intraclass correlation
CV = coefficient of variation
WNS = within-subject means square
SE = standard error

effect: as the study progresses, the staff gets better at taking measurements. Temporary changes in error could be caused by equipment failure, the loss of key personnel, or even changes in local conditions (e.g., disruptions caused by building renovations). Over the course of time, these effects may be confounded and compounded as equipment wears out and breaks and as research personnel change. Even with the best quality assurance procedures in place, study drift can be a significant problem, especially in a multicenter study. This factor is of particular concern to the HERITAGE Family Study because subjects are recruited in batches over a 5-year period, with each subject measured twice, once before and once after a 20-week exercise training program.

Study drift can impact either type of error. Drift in bias may show up through mean measurements changing over time, and it is often easier to detect than bias itself. This is because detecting bias requires some indicator of the "true values," while detecting drift in bias merely requires comparing mean measurements over time (which could all be biased). While drift in bias should be avoided, it is sometimes possible to correct via a procedure such as normalizing the data by time period. Drift in measurement error, however, is more serious, particularly when the reliability of a measurement declines over time (i.e., the measurement error increases), because it can seldom be corrected. The overall repeatability of a measurement is dependent on both the precision of the measurement (how well repeat measurements made in a short time under the same conditions agree) and the biological variability in the measurement over time (e.g., a woman's progesterone level changes with the phase of her menstrual cycle, whereas blood pressure varies more unpredictably as a result of physical exertion, stress, and other factors). A measurement taken with a great deal of precision in each session might have poor reproducibility as a result of high variability between sessions, and the converse holds as well. Significant changes in either precision or variability over time would be a serious problem for a family study as the magnitude of family resemblance could be compromised simply due to changing reproducibility.

Although the overall reproducibility might be good, and most time periods might show little change in bias, there could be a time period when the reproducibility is significantly lower or the bias is significantly different. Such a period could contaminate the data and either reduce the power to detect effects or even produce spurious results. It is possible that eliminating all data from a certain period with much lower reproducibility could increase the power to detect effects. If bias changes over time, it would tend to distort the between-family variation while having little impact on the within-family variation because family members were tested at approximately the same time. If ignored, this bias drift would tend to distort the overall familial correlations, possibly even spuriously inferring familial aggregation. Thus it is important to identify if either type of drift occurs.

MATERIALS AND METHODS

Subjects and Experimental Design

Each subject in the HERITAGE Family Study was sedentary for at least 3 months before the study and then was put through a 20-week exercise training program consisting of three sessions each week on a cycle ergometer. To ascertain the effect of the exercise program, a large number of tests were performed both before training and after training. These included exercise tests (at maximal power output 50 watts power output, and 60% and 80% of the oxygen uptake in the maximal output test); anthropometric measurement (height, weight, skinfolds, etc.); body composition derived from underwater weighing; resting blood pressure; steroid hormone levels; lipid and lipoprotein levels; an intravenous glucose tolerance test; a dietary questionnaire; and a heparin clearance test. Details and a complete list, of tests are given in reference 1. Families were recruited, tested, trained, and retested, with white families consisting of two parents and at least three adult offspring, and black families consisting of at least two first-degree relatives. The study protocol was approved by each Clinical Center International Review Board (IRB), and informed consent was obtained from each subject.

As with any large-scale multicenter study, a number of methods were employed to ensure the quality of the data (3). For example, all blood assays were performed at central laboratories in Quebec, so as to avoid interlaboratory variation. Many measurements were repeated on different days, and most were collected more than once, with the intention of using the mean of the multiple measurements to reduce the measurement error. The steroid and lipid tests were performed in a 12-hour fasted state, and the women were in the early follicular phase of ovulation (thus reducing the effect of natural variation). The equipment used was also standardized across the centers, and the personnel were centrally trained and certified. To monitor and improve the data quality on an ongoing basis, each Clinical Center periodically recruited additional non-HERITAGE volunteers, on whom key components of the protocol were administered multiple times.

For this analysis, the pretraining data on all subjects who

AEP Vol. 7, No. 7
October 1997: 452–462

TABLE 1. Demographics by period for 50-watt exercise tests

Period	N	Male (%)	Black (%)	Age (years)		
				Mean	Minimum	Maximum
1. January–March 1993	72	47	14	37	17	63
2. April–July 1993	65	51	9	34	17	59
3. August–November 1993	59	44	25	35	17	64
4. December 1993–March 1994	64	59	14	36	17	60
5. April–July 1994	70	49	24	35	17	65
6. August–December 1994	57	54	49	33	17	61

had completed training (n = 336) or dropped out (n = 53; three for injury or illness, five for pregnancy, 17 refused further study, 19 for poor compliance, and 9 because other family members dropped out) as of March 25, 1995, were used (which represents 60% of the total recruitment goal of 650). These subjects represent the HERITAGE study master data set version 1.10, which has been extensively checked for data entry errors and consistency. The measurements examined here are from the 50-watt power output level during the exercise test, the anthropometric measurements, the resting blood pressures, and the lipid and lipoprotein assays. Since an extensive analysis of the exercise test data (J.H. Wilmore, P.R. Stanforth, K.R. Turley, et al., unpublished data) indicates that the measurements at 50 watts are the least reproducible of the exercise test data, they were considered here to examine drift over time. Details of all measurements, tests, and assays are given in the HERITAGE Manual of Procedures (4).

The 50-watt exercise test was administered twice on separate days. The subject maintained a constant 50-watt power output for 8–12 minutes while on a cycle ergometer, during which time the measurements were obtained twice, yielding four measurements of each variable (two on each day). A number of variables were measured, and here we looked at eight of them: systolic blood pressure, diastolic blood pressure, heart rate, expired volume, carbon dioxide production, oxygen uptake, cardiac output, and stroke volume.

For each of the anthropometric variables, two valid measurements were obtained in a single session (4). Essentially, if the first two measurements were not within a certain prespecified tolerance, a third measurement was taken. The two measurements closest to each other were taken as the valid measurements. Eight of these variables were examined here: height, weight, waist circumference, upper arm length, and skinfolds at the abdomen, biceps, calf, and subscapular region.

Six valid resting blood pressure measurements were recorded on each subject, with three measurements obtained on each of 2 days (4). On each of the 2 days, the first reading was deleted. From the remaining, a reading was considered valid if the manual and automated (COLIN) readings were within a prespecified tolerance. A total of three valid readings were sought on each day. Diastolic blood pressure, systolic blood pressure, and heart rate were all examined here.

The lipid assays were done twice, with blood samples drawn at least 24 hours apart. The blood was drawn at each Clinical Center and prepared according to a standard protocol before being sent to the central laboratory in Quebec for analysis. Eleven of these variables were selected for analysis here: plasma cholesterol, plasma triglycerides, low-density-lipoprotein (LDL) cholesterol, very-LDL (VLDL) cholesterol, high-density lipoprotein (HDL) cholesterol, HDL2 cholesterol, HDL3 cholesterol, apoprotein A1, total apoprotein B, and LDL apoprotein B. To eliminate subjects who did not follow the required fasting protocol, samples containing chylomicrons were not used in this analysis.

To assess the reproducibility over time, all subjects were divided into six 4-month time periods with roughly equal Ns (see Table 1): (*I*) January–March 1993; (*II*) April–July 1993; (*III*) August–November 1993; (*IV*) December 1993–March 1994; (*V*) April–July 1994; and (*VI*) August–December 1994. The first period is only three months long because no measurements were made in December 1992. The last period is slightly over four months long because master data set version 1.10 contained a few measurements made in December 1994, which were included in the final time period. Overall, of the 390 subjects there were 198 men, 192 women, 86 blacks, and 304 whites. The mean age was 34.9 (SD, 14.3; range 17–65 years).

Statistical Methods

From the repeated measurements, the reproducibility may be estimated using a linear gaussian model, such as an analysis of variance (ANOVA), which partitions the variance due to the different error sources (5). The error can then be quantified on both absolute and relative scales, with each having its advantages and disadvantages. Reported here are the traditional estimate on the absolute scale, namely the SD of the error effect, also called "technical error" (or within-subject SD); and two relative scale measures of reproducibility, the intraclass correlation coefficient and the within-subject coefficient of variation. The technical error (TE) gives direct information on the reproducibility of a measurement when the units and actual measurement values are familiar to the investigator (such as a TE of 0.5 cm for height). However, when the units or scale are not so familiar,

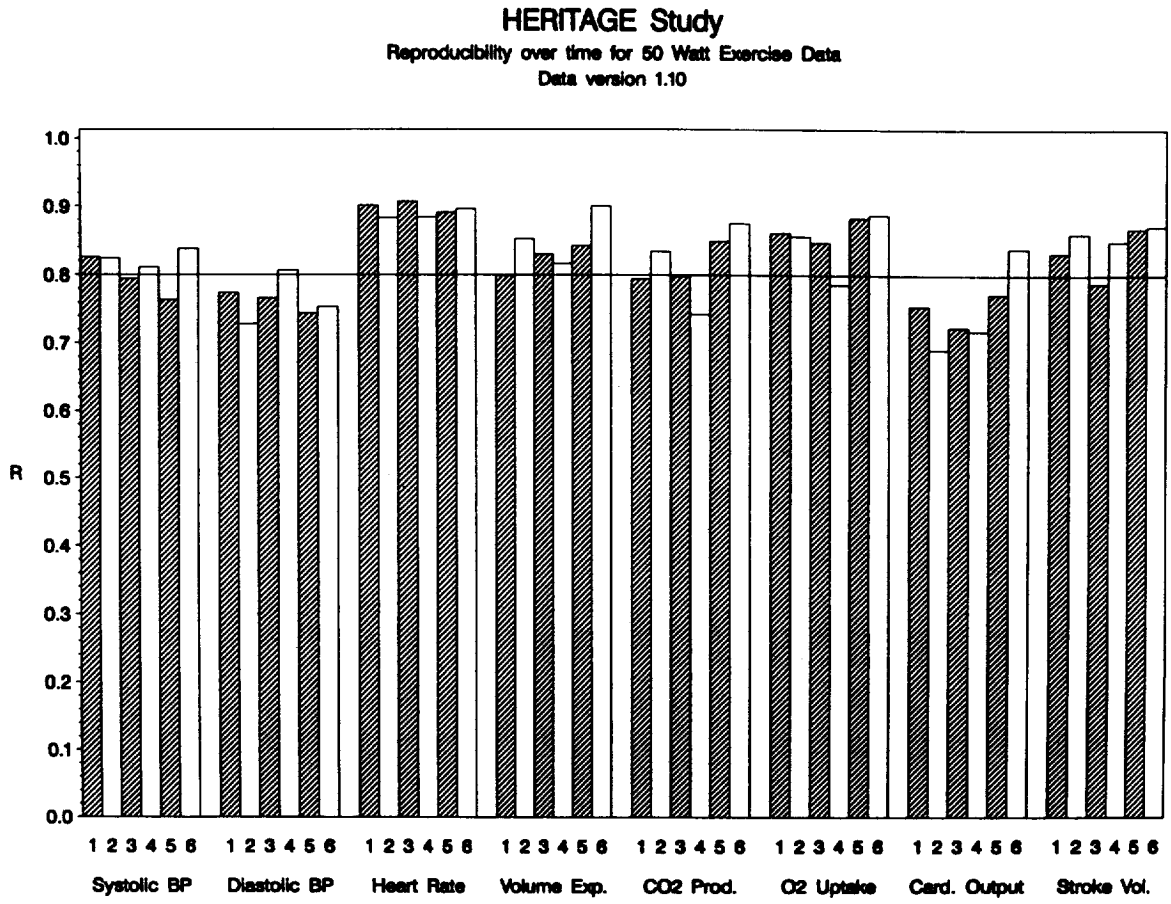

FIGURE 1. *ICCs* over time for systolic and diastolic blood pressure, heart rate, expired volume, carbon dioxide production, oxygen uptake, cardiac output, and stroke volume from the 50-watt exercise data.

TE may not be very meaningful. Furthermore, absolute measures such as TE are difficult to compare across phenotypes (e.g., which is more reproducible: height or cholesterol level?). Relative measures of reproducibility, such as intraclass correlation (*ICC*) and within-subject coefficient of variation (CV), automatically factor in estimates of the population values and are therefore easier to compare across scales. Thus, an *ICC* of > 0.8 or a CV of $< 10\%$ could be considered as indicators of very good reproducibility. However, the relative nature of the scale can be misleading, especially if the underlying population variation is small. For example, if measuring height of people over 6 feet tall, people under 6 feet tall, and people exactly 6 feet tall, the TE for each group would probably be the same as the TE for the combined group, but the *ICCs* for the over 6 feet and the under 6 feet groups would be smaller than the *ICC* for the combined group, and the *ICC* for the exactly 6 feet group would be almost zero. For the *ICC* to be valid and meaningful, it is important that there be enough variability among the subjects.

The *ICC*, TE, and CV were all computed under the model (1,1) of Shrout and Fleiss (5). Under this model, the i^{th} measurement on the j^{th} subject, x_{ij}, is given by:

$$x_{ij} = \mu + b_j + \omega_{ij},$$

where μ is the population mean b_j is the j^{th} subject effect, and ω_{ij} is the residual. Both b_j and ω_{ij} are assumed to be normally distributed and independent with SDs of σ_τ and σ_ω, respectively. Whereas the TE is σ_ω, the CV is computed as:

$$CV = (100 * \sigma_\omega)/\mu.$$

The standard formula for the *ICC* is:

$$ICC = \sigma_\tau^2/(\sigma_\tau^2 + \sigma_\omega^2).$$

To compute the *ICC*, PROC GLM in SAS was used to run an ANOVA, yielding a between-subjects mean square (*BMS*) and within-subjects mean square (*WMS*). *WMS* is an unbiased estimate of σ_ω^2, and σ_τ^2 may be estimated by $(BMS-WMS)/k$, where k is the number of repeat measurements on a subject. This yields an estimate of the *ICC* as:

$$ICC = (BMS - WMS)/[BMS + (k - 1)WMS].$$

AEP Vol. 7, No. 7
October 1997: 452–462

TABLE 2. Overall mean, technical error (TE), coefficient of variation (CV), and intraclass correlation (ICC) over six time periods for 50-watt exercise variables

	Period					
Variable[a]	1 (N = 72)	2 (N = 65)	3 (N = 59)	4 (N = 64)	5 (N = 70)	6 (N = 57)
Systolic BP (50 W) (mm Hg)						
Mean	149.55	144.81	146.22	147.40	143.41	151.62
TE	9.13	8.79	9.67	9.40	9.30	8.69
CV	6.14	6.10	6.66	6.39	6.49	5.72
ICC	0.83	0.82	0.79	0.81	0.76	0.84
Diastolic BP (50 W) (mm Hg)						
Mean	74.00	69.97	74.19	74.30	72.15	75.64
TE	5.95	6.52	6.56	5.68	6.30	6.77
CV	8.09	9.34	8.90	7.65	8.74	8.97
ICC	0.77	0.73	0.77	0.81	0.74	0.75
Heart rate (50 W) (bmp)						
Mean	125.57 (5)[b]	117.95	124.10 (5)	117.11	113.69 (1,3)	119.96[c]
TE	6.60	6.35	5.94	5.95	5.33	5.85
CV	5.27	5.37	4.79	5.09	4.69	4.89
ICC	0.90	0.88	0.91	0.88	0.89	0.90
Expired volume (l/min)						
Mean	34.40 (3,5)	31.87	31.07 (1)	31.72	29.32 (1)	31.63[c]
TE	2.38	1.96	2.11	2.45	2.13	2.32
CV	7.00	6.17	6.81	7.70	7.26	7.35
ICC	0.80	0.85	0.83	0.82	0.84	0.90
Carbon dioxide production (ml/min)						
Mean	1055.25[d]	991.41 (1,5)	943.01 (1)	992.88 (1,5)	910.73 (1,2,4)	946.33 (1)[c]
TE	52.88	45.68	47.87	66.10	43.03	48.43
CV	5.03	4.61	5.08	6.64	4.72	5.12
ICC	0.79	0.84	0.80	0.74	0.85	0.87
Oxygen uptake (ml/min)						
Mean	1125.21 (2,3,5,6)	1066.26 (1,5)	1016.42 (1,4)	1094.18 (3,5,6)	992.07 (1,2,4)	1019.64 (1,4)[c]
TE	45.20	48.62	41.61	61.60	39.84	41.28
CV	4.02	4.56	4.09	5.62	4.02	4.04
ICC	0.86	0.86	0.85	0.79	0.88	0.89
Cardiac output (l/min)						
Mean	12.08 (3,5,6)	11.76 (6)	11.25 (1)	11.67	11.00 (1)	10.85 (1,2)[c]
TE	0.97	0.93	0.96	0.86	0.75	0.63
CV	7.95	7.90	8.55	7.43	6.82	5.83
ICC	0.76	0.69	0.72	0.72	0.77	0.84
Stroke volume (ml/beat)						
Mean	98.15	103.32 (3,6)	92.24 (2)	102.37	98.68	92.02 (2)[c]
TE	9.02	8.77	9.12	7.65	7.41	6.69
CV	8.97	8.57	9.82	7.47	7.51	7.26
ICC	0.83	0.86	0.79	0.85	0.87	0.87

[a] Abbreviations: BP, blood pressure; W, watt.
[b] Fails the homogeneity test and has significantly different pairs of periods.
[c] Numbers in parentheses indicate study periods that have a significantly different value.
[d] All other periods have a significantly different value.

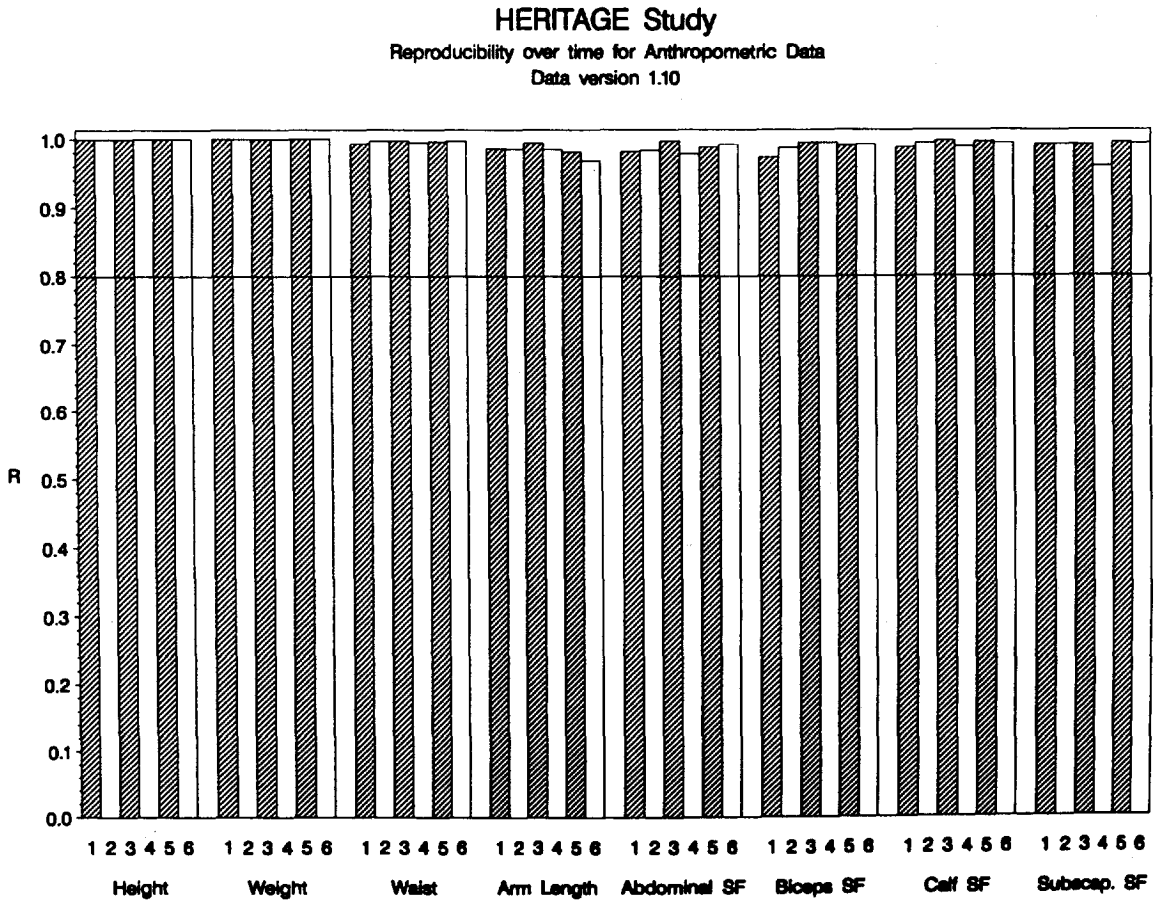

FIGURE 2. *ICCs* over time for height, weight, waist circumference, upper-arm length, and skinfolds at the abdomen, biceps, calf, and subscapular area from the anthropometric data.

Using the *ICC* and an approximate SE for each of the six time periods, a χ^2-square test of homogeneity was computed for each variable that had an *ICC* of < 0.95 for any period. Since any *ICC* over 0.95 is exceptionally good, any difference between *ICCs* in this range is almost meaningless and was therefore not pursued. If this homogeneity χ^2-square was found to be significant, approximate z-scores were used to determine whether the *ICCs* for any pair of periods were significantly different:

$$z_{kl} = (|ICC_k - ICC_l|) \Big/ \sqrt{\overline{ICC} \cdot (1 - \overline{ICC}) \cdot \left(\frac{1}{n_k} + \frac{1}{n_l}\right)}$$

where ICC_k is the *ICC* for time period k, ICC_l is the *ICC* for time period l, n_k and n_l are the number of observations in the respective time periods, and $\overline{ICC}$ is a weighted average of the two *ICCs*.

The *ICCs*, *CVs*, and TEs were computed here from individual measurements. However, it was planned that averages of these measurements would be used later in analysis of the effects of the exercise program. Since using the average of several measurements generally improves reproducibility, the values reported here probably underestimate the reproducibility of the final data set.

Drift in the mean was assessed from the average of the multiple measurements. To test if there were differences among the six time periods, a linear model was implemented using PROC GLM in SAS. If a significant difference was found, Tukey's studentized range test, which corrects for multiple comparisons, was used to determine between which periods there were significant differences. It is important to clarify that most of the variables examined here are positively correlated within families. Using family data in such a situation is known to underestimate the SEs, and therefore to inflate the significance of differences. Hence, as a result of this alone, one should expect to see a few spurious differences.

RESULTS

For each variable considered, the means, TEs, *CVs*, and *ICCs* for each of the six time periods are presented in Tables 2–5.

AEP Vol. 7, No. 7
October 1997: 452–462

TABLE 3. Overall mean, technical error (TE), coefficient of variation (CV), and intraclass correlation (*ICC*) over six time periods for anthropometric variables

	Period					
Variable	1 (*N* = 77)	2 (*N* = 60)	3 (*N* = 59)	4 (*N* = 52)	5 (*N* = 78)	6 (*N* = 60)
Height (cm)						
Mean	171.34	171.35	167.13 (5)[a]	171.13	172.27 (3)	170.43[b]
TE	0.11	0.14	0.17	0.15	0.17	0.15
CV	0.07	0.08	0.10	0.09	0.10	0.09
ICC	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]	
Weight (kg)						
Mean	79.14	78.30	70.91 (4)	81.45 (3)	74.41	73.11[b]
TE	0.32	0.05	0.08	0.05	0.05	0.10
CV	0.40	0.06	0.11	0.07	0.07	0.13
ICC	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]
Waist circumference (cm)						
Mean	93.92 (5,6)	91.49	87.54 (4)	97.41 (3,5,6)	86.86 (1,4)	84.67 (1,4)[b]
TE	0.44	0.44	0.40	0.46	0.41	0.38
CV	0.48	0.48	0.46	0.47	0.47	0.45
ICC	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]
Upper arm length (cm)						
Mean	36.12 (5)	36.56	36.35 (5)	37.21	37.58 (1,3)	36.86[b]
TE	0.29	0.27	0.19	0.27	0.26	0.30
CV	0.80	0.73	0.52	0.74	0.70	0.80
ICC	0.99	0.99	0.99	0.99	0.99	0.98
Abdomen skinfold (mm)						
Mean	27.05	27.50	22.89 (4)	29.40 (3,5,6)	22.99 (4)	22.64 (4)[b]
TE	0.46	0.62	0.39	0.55	0.44	0.46
CV	1.69	2.26	1.70	1.86	1.91	2.01
ICC	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]
Biceps skinfold (mm)						
Mean	10.00	10.11	8.09	10.73	8.48	8.74
TE	0.64	0.35	0.36	0.32	0.36	0.27
CV	6.38	3.47	4.49	2.99	4.22	3.04
ICC	0.99	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]
Calf skinfold (mm)						
Mean	17.56	19.09	15.54	17.40	16.14	14.41
TE	0.53	0.36	0.34	0.31	0.37	0.41
CV	3.03	1.89	2.21	1.80	2.27	2.85
ICC	0.99	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]
Subscapular area skinfold (mm)						
Mean	18.52	18.18	15.59 (4)	20.97 (3,5)	16.28 (4)	16.60[b]
TE	0.46	0.33	0.42	0.37	0.35	0.37
CV	2.47	1.80	2.67	1.76	2.17	2.23
ICC	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]	0.99[c]

[a] Numbers in parentheses indicate study periods that have a significantly different value.
[b] Fails the homogeneity test and has significantly different pairs of periods.
[c] Value over 0.995 truncated to 0.99.

The means and *ICC*s that fail the homogeneity test across the six periods are identified in the tables. For better visualization, charts of the *ICC*s are presented in Figures 1–4.

Nearly all of the *ICC*s indicate an excellent level of reproducibility. Only two of the 29 variables examined have even a single period in which the *ICC* is < 0.7. The eight anthropometric measurements, for example, indicate exceptional reproducibility, with the lowest ICC in this group of measurements being 0.98, while most are over 0.99.

The homogeneity of the *ICC*s for each variable follows a pattern similar to that of the overall quality of the *ICC*s for that variable: Only one of the variables examined was found to have significant differences in *ICC*s between periods, and it is the resting heart rate, which has a period with an *ICC* < 0.7. However, the primary difference for resting heart rate is a smaller TE in period 2. In addition, the *ICC*s for apo-A1 failed the homogeneity test, though the post-hoc tests found no differences between periods.

The CVs are also generally good, with only four variables having periods with a CV over 10%. VLDL-cholesterol has

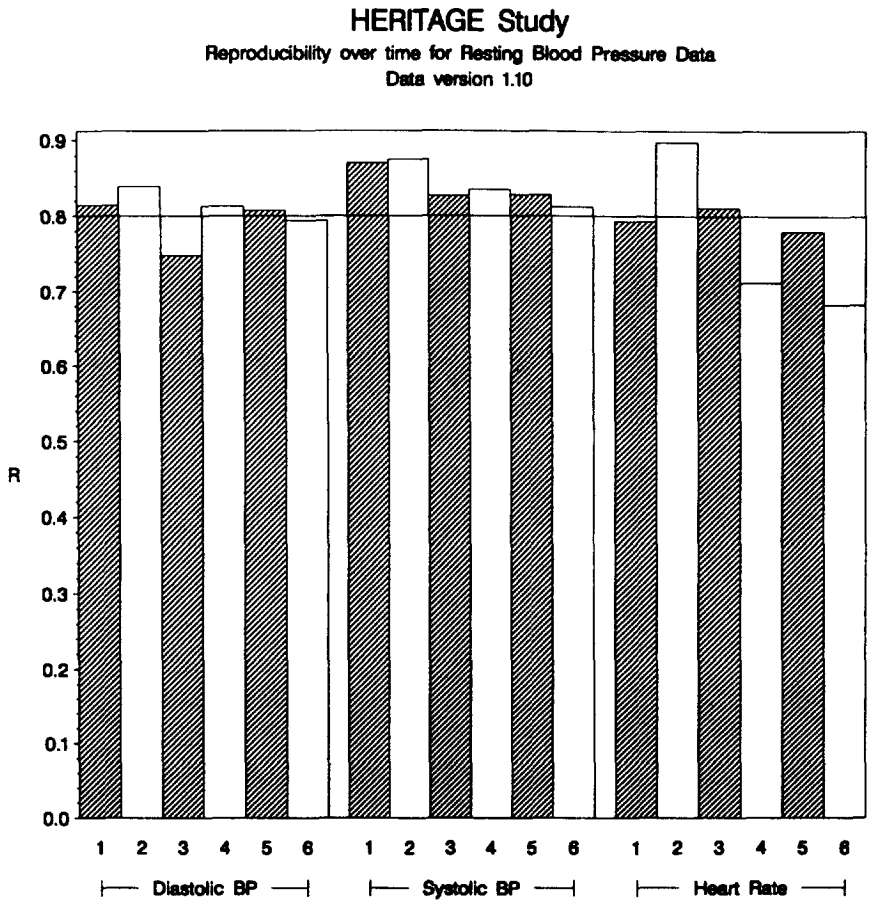

FIGURE 3. *ICCs* over time for diastolic and systolic blood pressure and heart rate from the resting blood pressure data.

CVs consistently over 20%, with one value near 40%. Triglycerides have *CVs* from slightly below 20% to slightly over 30%. HDL2-cholesterol also has high *CVs*, with all values between 10% and 20%. Only one other test has even a single *CV* over 10%: HDL3-cholesterol in period 2 at 10.75%.

To provide a further comparison of the six time periods, for each variable the periods with the worst and next-to-worst *ICC* and *CV* were examined. We found that the 50-watt exercise data, anthropometric data, and resting blood pressure data show no discernible concentration of worst or low *ICCs*. For the lipid tests, period 6 seems to suggest a worsening *ICC*. However, additional data collected since completion of the data analyses reported here suggest that this was a sporadic case and that there was no declining trend in the *ICCs*.

The means of each variable across the six periods are also shown in Tables 2–5. Six of the exercise tests are found to have significant differences in the mean values over the six periods. The post-hoc Tukey tests found 25 pairs of periods across these exercise tests that are significantly different. It should be noted that period 1 is included in 15 of these 25 pairs, suggesting that it is different from the others. Six of the eight anthropometric measurements show at least one significant difference. For the resting blood pressure, only periods 2 and 6 for systolic blood pressure are significantly different. Three of the eleven lipid tests have *P* values below 5% in the test for differences across periods, but, in one of these (HDL3-cholesterol), the Tukey test finds no pair of periods that are significantly different from each other.

Before the results reported here were obtained, the data were thoroughly inspected, which improved the overall quality. In particular, we found chylomicrons were present in the lipid profiles of four subjects, indicating that they did not fast before the lipid tests. Including these four samples would have reduced several of the *ICCs*. For example, the most striking effect was in triglycerides in period 6, where the *ICC* would have been reduced to 0.61 from the 0.74 reported here.

DISCUSSION

The major goal of this investigation was to determine whether any study drift existed in this multicenter family exercise study (HERITAGE) where the data collection ex-

AEP Vol. 7, No. 7
October 1997: 452–462

TABLE 4. Overall mean, technical error (TE), coefficient of variation (CV), and intraclass correlation (ICC) over six time periods for resting blood pressure

	Period					
Variable[a]	1 (*N* = 66)	2 (*N* = 67)	3 (*N* = 60)	4 (*N* = 52)	5 (*N* = 74)	6 (*N* = 62)
Diastolic blood pressure (rest, mm Hg)						
Mean	68.36	65.06	66.54	68.33	67.27	68.89
TE	4.30	4.31	4.37	4.21	4.13	3.94
CV	6.29	6.61	6.57	6.18	6.14	5.73
ICC	0.81	0.84	0.75	0.81	0.81	0.79
Systolic blood pressure (rest, mm Hg)						
Mean	118.70	115.22 (6)[a]	117.02	119.39	118.31	121.76 (2)[b]
TE	4.52	4.28	4.96	5.20	4.42	4.57
CV	3.81	3.71	4.24	4.36	3.73	3.75
ICC	0.87	0.87	0.83	0.84	0.83	0.81
Heart Rate (rest, beats/min)						
Mean	65.83	64.56	65.01	63.90	64.68	66.11
TE	4.55	3.63	4.15	4.18	3.92	4.66
CV	6.91	5.62	6.39	6.55	6.04	7.07
ICC	0.79	0.90 (4,6)	0.81	0.71 (2)	0.78	0.68 (2)[b]

[a] Numbers in parentheses indicate study periods that have a significantly different value.
[b] Fails the homogeneity test and has significantly different pairs of periods.

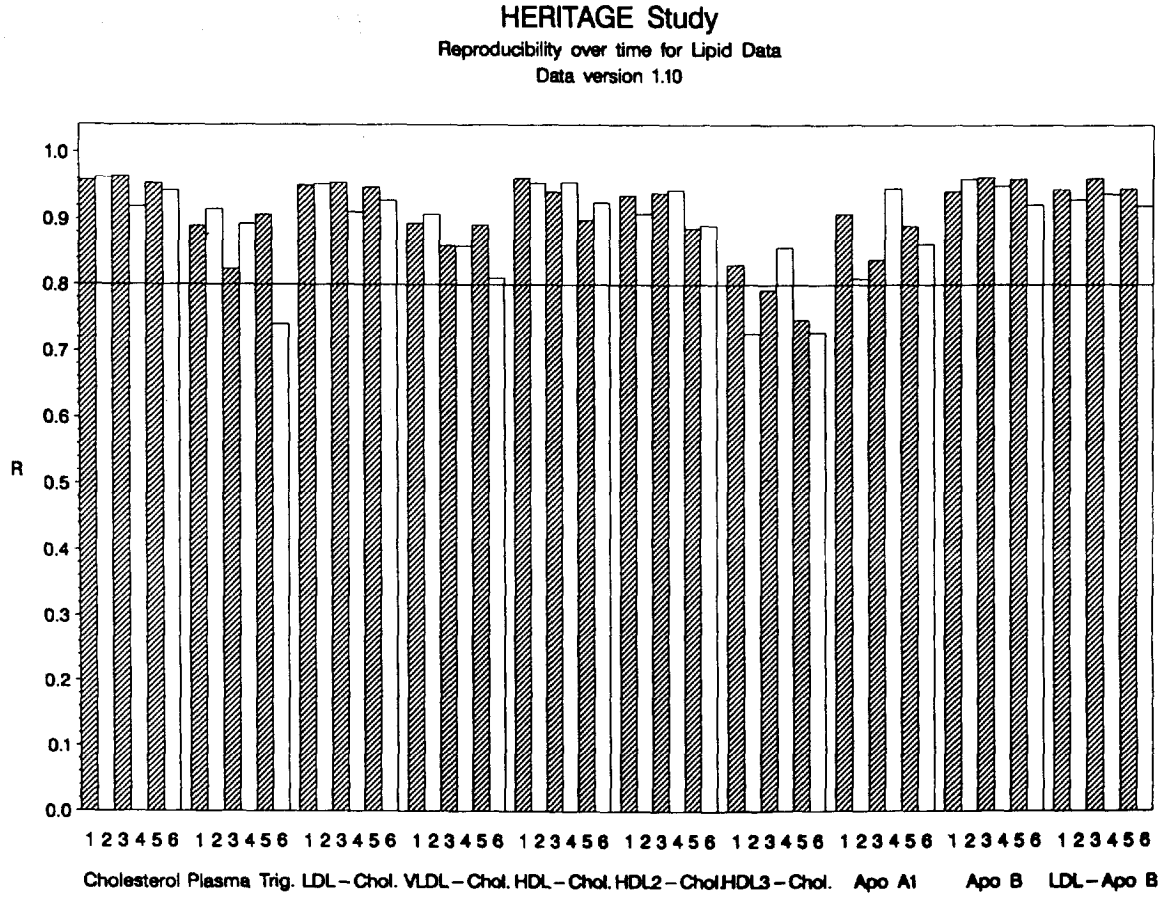

FIGURE 4. ICCs over time for cholesterol, plasma triglycerides, LDL-cholesterol, VLDL-cholesterol, HDL-cholesterol, HDL2-cholesterol, HDL3-cholesterol, apoprotein A1, total apoprotein B, and LDL-apoprotein B, from the lipid data.

TABLE 5. Overall mean, technical error (TE), coefficient of variation (CV), and intraclass correlation (ICC) over six time periods for lipid variables

	Period					
Variable	1 (*N* = 66)	2 (*N* = 65)	3 (*N* = 59)	4 (*N* = 52)	5 (*N* = 77)	6 (*N* = 59)
Cholesterol (mmol/L)						
Mean	4.46	4.52	4.66	4.33	4.40	4.35
TE	0.19	0.18	0.23	0.26	0.19	0.22
CV	4.18	4.04	4.94	6.01	4.42	4.94
ICC	0.96	0.96	0.96	0.92	0.95	0.94
Plasma triglycerides (mmol/L)						
Mean	1.28	1.38	1.31	1.31	1.29	1.23
TE	0.23	0.25	0.41	0.25	0.28	0.32
CV	18.19	18.09	31.69	19.83	22.29	26.15
ICC	0.89	0.91	0.82	0.89	0.91	0.74
LDL-cholesterol (mmol/L)						
Mean	3.07	3.03	3.15	2.93	2.88	2.92
TE	0.19	0.18	0.21	0.24	0.17	0.20
CV	6.04	5.98	6.70	8.20	5.81	6.75
ICC	0.95	0.95	0.95	0.91	0.95	0.93
VLDL-cholesterol (mmol/L)						
Mean	0.39	0.42	0.44	0.43	0.43	0.41
TE	0.09	0.11	0.17	0.11	0.14	0.13
CV	23.09	26.93	39.91	26.36	34.18	31.17
ICC	0.89	0.91	0.86	0.86	0.89	0.81
HDL-cholesterol (mmol/L)						
Mean	1.00	1.07	1.07	0.98	1.09	1.03
TE	0.05	0.06	0.06	0.07	0.09	0.06
CV	4.69	5.59	5.27	7.19	7.83	5.44
ICC	0.96	0.95	0.94	0.96	0.90	0.93
HDL2-cholesterol (mmol/L)						
Mean	0.31 (2)[a]	0.41 (1,6)	0.38	0.32	0.38	0.31 (2)[b]
TE	0.04	0.06	0.05	0.05	0.07	0.05
CV	12.58	15.80	12.07	15.99	18.43	15.68
ICC	0.94	0.91	0.94	0.94	0.89	0.89
HDL3-cholesterol (mmol/L)						
Mean	0.69	0.66	0.70	0.66	0.72	0.72[c]
TE	0.06	0.07	0.05	0.06	0.07	0.06
CV	8.08	10.75	7.66	9.31	9.64	8.83
ICC	0.83	0.73	0.79	0.86	0.75	0.73
Apolipoprotein A1 (g/L)						
Mean	1.10 (5,6)	1.15	1.17	1.10 (5,6)	1.22 (1,4)	1.23 (1,4)[b]
TE	0.05	0.07	0.06	0.05	0.05	0.06
CV	4.27	6.04	5.13	4.66	4.48	4.51
ICC	0.91	0.81	0.84	0.95	0.89	0.86[c]
Tot. Apolipoprotein B (g/L)						
Mean	0.85	0.86	0.90	0.85	0.82	0.86
TE	0.06	0.04	0.05	0.05	0.05	0.06
CV	6.96	5.19	5.74	6.23	5.67	7.13
ICC	0.94	0.96	0.96	0.95	0.96	0.92
LDL-Apolipoprotein B (g/L)						
Mean	0.76	0.77	0.81	0.78	0.75	0.80
TE	0.05	0.05	0.05	0.05	0.05	0.06
CV	6.73	6.57	5.73	6.82	6.51	7.08
ICC	0.94	0.93	0.96	0.94	0.95	0.92

[a] Periods which have a significantly different value.
[b] Fails the homogeneity test and has significantly different pairs of periods.
[c] Fails the homogeneity test.

AEP Vol. 7, No. 7
October 1997: 452–462

tended over 5 years. Unfortunately, this aspect has not been well studied in the past, perhaps in part because there have been few exercise training studies of a large scope in both the study duration and the number of subjects.

The results for the exercise, lipid, anthropometric, and blood pressure data show that these HERITAGE data were, for the most part, highly reproducible. Not only was the overall reproducibility for these tests good, but it is consistently good over time. In particular, there was no pattern suggesting decreasing reproducibility over time. Of the 29 variables examined, only one (resting heart rate) showed any significant reproducibility differences over time.

The significant reproducibility changes in the resting heart rate may have resulted from random effects that were due in part to the large number of tests done here (even with some correction for multiple comparison). The resting heart rate is influenced by a number of factors, and it is sometimes difficult to get subjects into a truly relaxed and undisturbed state. However, the reproducibility was fairly acceptable most of the time periods.

We have focused primarily on the *ICC* rather than the *CV* because *CV*s are corrected for the mean value of a measurement while *ICC*s are corrected for the variance. Thus the *ICC*s give a better indication of the predictive value of a measurement, i.e., how useful the measurement is in distinguishing different individuals or states. However, most of the *CV*s were quite good. Although four of the lipid variables have *CV*s over 10%, their *ICC*s are relatively good, indicating that the magnitude of these variables was small relative to the population variance, rather than any problem with the quality of the measurements.

Most of the significant differences in mean are probably due to gender and between-family variation rather than to systematic measurement error. In part, this was because the tests for drift in bias were not corrected for ascertainment of the data in families. Furthermore, the male/female ratio changed from period to period, with the lowest ratio in period 3 and the highest in period 4. There were a large number of significant differences in the anthropometric measurements, but these measurements had exceptional ICCs. Period 3 had the lowest mean weight and the lowest male:female ratio, while period 4 had the highest mean weight and the highest male:female ratio. Indeed, when weight was considered separately for men and women, no significant differences between periods were found.

With the exercise tests, an added complication was that some of the measures vary with body size. While one might suspect some sort of training effect for these tests since 15 of the 25 pairs of periods with significant differences in the exercise test data included the first period, closer examination showed that for expired volume, carbon dioxide production, oxygen uptake, cardiac output, and stroke volume, the three largest values occurred in periods 1,2, and 4. These periods corresponded exactly with the tree periods with the highest mean body weight.

In summary, most measurements in the HERITAGE family study were highly reproducible. There was little evidence to suggest drift over time in either the reproducibilities or the mean values.

The HERITAGE Family Study is supported by the National Heart, Lung and Blood Institute through the following grants: HL45670 (C. Bouchard); HL47323 (A.S. Leon); HL47317 (D.C. Rao); HL47327 (J.S. Skinner); and HL47321 (J.H. Wilmore). Credit is also given to the University of Minnesota Clinical Research Center, NIH Grant MO1-RR000400. Thanks are expressed to all of the co-principal investigators, investigators, co-investigators, local project coordinators, research assistants, laboratory technicians, and secretaries who have contributed to this study. Gratitude is also expressed to the members of the Advisory Board. Finally, the HERITAGE consortium is very thankful to those hard-working families whose participation has made these data possible.

REFERENCES

1. Bouchard C, Leon AS, Rao DC, Skinner JS, Wilmore JH, Gagnon J. The HERITAGE Family Study: Aims, design, and measurement protocol. Med Sci Sports Exerc. 1995;27:721–729.
2. Sievänen H, Oja P, Vuori I. Scanner-induced variability and quality assurance in longitudinal dual-energy x-ray absorptiometry measurements. Med Phys. 1994;21:1795–1805.
3. Gagnon J, Province MA, Bouchard C, et al. The HERITAGE Family Study: Quality assurance and quality control. Ann Epidemiol. 1996; 6:520–529.
4. HERITAGE Consortium Steering Committee. The HERITAGE Family Study: Manual of Procedures. Laval University, Quebec, QC: Physical Activity Sciences Laboratory; 1996.
5. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86:420–428.