Reliability of hormonal levels for assessing the hypothalamic-pituitary-adrenocortical system in clinical pharmacology

J. COSTE,[1,2], G. STRAUCH[2,3], M. LETRAIT[2] & X. BERTAGNA[3]

[1]Département de Biostatistique et d'Informatique Médicale, [2]ECLIMED-Institut de Recherche Thérapeutique, and [3]Service d'Endocrinologie et Métabolisme, Hôpital Cochin, 27, rue du Faubourg Saint-Jacques, 75674 Paris Cedex 14, France

Few data are available on the reliability of measurements of adrenocortical and corticotroph hormones for use in clinical pharmacology. Two placebo controlled crossover trials in 20 normal healthy male subjects offered the opportunity to perform three repeat samplings of adrenocortical and corticotroph hormones at 1 to 5 week intervals during the placebo periods. Measurements of baseline levels of plasma, salivary and urinary cortisol, plasma adrenocorticotroph hormone (ACTH), lipotrophic hormone (LPH), β-endorphin, post tetracosactrin levels of plasma and salivary cortisol, post corticotrophin releasing hormone (CRH)-lysine vasopressine (LVP) levels of plasma cortisol, ACTH and LPH; and post metyrapone levels of plasma cortisol and 11-deoxycortisol (compound S), ACTH, LPH, β–endorphin were performed in the same laboratory. The reliability of the measurements was estimated by computing the intraclass correlation coefficient (*R*) and by using Altman-Bland graphical method. The *R*s of baseline parameters varied from 0.18 (for 08.00 h salivary cortisol) to 0.55 (for 08.00 h plasma cortisol and nocturnal urinary cortisol). In contrast, parameters obtained after direct stimulation or inhibition of the producing targets were much more reliable: *R*s were above 0.80 for post tetracosactrin levels of plasma and salivary cortisol, post CRH-LVP levels of plasma ACTH and LPH. The *R*s were below 0.50 for post metyrapone levels of plasma 11-deoxycortisol, ACTH, LPH and β-endorphin. The interval between sampling did not affect *R* estimates. These data show that peak levels of plasma cortisol and ACTH after direct stimulation are highly reliable whereas baseline and main post-metyrapone levels are not. A single post stimulation level of these hormones may suffice to characterize normal subjects in trials while three or more baseline or post-metyrapone level measurements may be needed to obtain a satisfactory reliability of the mean.

Keywords hormonal levels pituitary-adrenal axis reliability

Introduction

The hypothalamic-pituitary-adrenocortical (HPA) system has three main functional characteristics: i) basal activity leading to appropriate levels of cortisol production in a circadian pattern, ii) negative feedback regulation that results in stimulation of the system when plasma cortisol levels fall and iii) responses to a variety of stresses which, if strong enough, will override the circadian and feedback controls [1]. Each or all of these functional characteristics may be altered by several classes of drugs, and such alterations can have serious consequences [1]. Methods for comprehensively assessing the functioning of the HPA system are therefore necessary during the development of such drugs. In clinical pharmacology, the

Correspondence: Dr J. Coste, Département de Biostatistique et d'Informatique Médicale, Hôpital Cochin, 27, rue du Faubourg Saint-Jacques, 75674 Paris Cedex 14, France

availability of powerful screening procedures is essential to support or discard suspicions of induced dysfunction. Obviously, reliable and valid procedures are required. In particular, the ability of a method to evidence moderate or mild dysfunctioning depends in part on the variability of the measurement. The variability of a series of measurements performed on different subjects using the same protocol (especially at the same moment when a circadian pattern is suspected) can be broken down into two elements: the *biological variability* (between-person variability) and the *random error variability* (which includes both within-person variability and method error). A classical index of reliability is the *intraclass correlation coefficient* which is typically the ratio of the variance due to the biological variability over the sum of the variance due to biologic variability and the variance due to random error variability [2].

Whereas the validity [3] and the method error [1] of most of the HPA system circulating hormones have been assessed for diagnostic purposes, between-person variability and within-person variability have not been systematicaly examined. Two placebo controlled cross-over trials in healthy male subjects offered the opportunity to evaluate the reliability of the hormonal parameters commonly used to assess the normality of HPA function.

Methods

Subjects

The subjects were 20 normal male volunteers, ranging in age from 21–29 years, who gave informed and written consent to participate in two phase I trials (10 subjects were included in each trial). Physical examination, routine laboratory data and pituitary-adrenal axis investigation, including baseline plasma cortisol and tetracosactrin stimulation, showed no evidence of any endocrine or other disease immediately before and during the study. All had normal sleep-activity cycles and feeding habits, were within 15% of the ideal body weight, and were medication free during the study period.

Study design

The same double-blind randomized placebo cross-over format was used for the two studies: subjects received 8 days of drug (or placebo) in the first period (day 1 to day 8) followed by a 28-day washout period before resumption of treatment for the remaining 8 days (second period) with placebo (or drug). In both studies, the hormonal profile was evaluated on day 1 (before drug administration) and day 9 of each period. Since no carry-over effect was observed in the two studies, three hormonal profile evaluations for each subject were available for the reliability study: two measurements performed in the placebo period and one measurement performed before drug administration in the treatment period. (Therefore, no measurement during drug periods was used). The two studies were conducted over the same period (September–December 1991) after approval by the Cochin Hospital Institutional Review Board.

In the first study, the hormonal profile evaluation included i) a baseline assessment: 08.00 h plasma and salivary cortisol, nocturnal (22.00 h–08.00 h) urinary cortisol, 08.00 h plasma adrenocorticotroph hormone (ACTH) and lipotrophic hormone (LPH); ii) an evaluation of adrenocortical reserves: plasma and salivary cortisol peak after direct stimulation by synthetic ACTH (tetracosactrin 0.25 mg i.m.) and indirect stimulation (corticotrophin releasing hormone (CRH)-lysine vasopresine (LVP) test: 100 mg CRH i.v. followed by 1 i.u. LVP over 15 min infusion), iii) an evaluation of pituitary reserve: plasma ACTH, plasma LPH peak during direct stimulation (CRH-LVP test).

In the second study, the hormonal profile included i) a baseline assessment: 08.00 h plasma and 24-h urinary cortisol, 08.00 h plasma βendorphin, 08.00 h plasma ACTH and LPH, ii) an evaluation of the feedback control of HPA system after metyrapone: plasma 11-deoxycortisol (compound S) and cortisol, plasma ACTH, LPH, plasma βendorphin (short metyrapone test: single oral administration of metyrapone, 30 mg kg^{-1}, at 24.00 h with measurement of hormones 8 h later).

Laboratory methods

Blood (or saliva/urine) was collected three times from each subject, with a 1- and 5-week interval between each collection. Cells were separated from plasma by centrifugation and removed. Aliquots of plasma were prepared from each subject and stored at –70°C. Plasma, salivary and urinary cortisol, plasma compound S and plasma LPH were determined using assays as previously described [4, 5, 6]. Plasma ACTH was measured by the ELSA-ACTH-immunoradiometric assay (CIS-Bio International, France). Plasma β-endorphin was measured by the Allegro immunoradiometric assay (Nichols Institute, California, USA). All samples from an individual subject were analysed in the same assay in the same laboratory. The within- and between-assay variation were as follows: cortisol 5 and 12%, ACTH 3 and 5%, LPH 3 and 5%, β-endorphin 4 and 9%.

Statistical methods

The statistical methods described by Fleiss [7] for the analysis of the reliability of quantitative data were used. Intraclass correlation coefficients (*R*) and their 95% confidence intervals were computed [8]. The intraclass correlation coefficient expresses the relative magnitude of the two components of the total variability i.e. the biological variability (between person variability: σ^2_{BP}) and random error σ^2_{error}, which includes both within-person variability and method error), in a series of measurements on different subjects.

$R = \dfrac{\sigma^2_{BP}}{\sigma^2_{BP} + \sigma^2_{error}}$ can be interpreted as the correlation coefficient between repeated measurement for a person. R values close to 1 ($R > 0.8$) indicate that a single measurement would be satisfatory to classify an individual with respect to the analyte. For less reliable measurements (when $R \leq 0.8$), we calculated the minimum number of replicate measurements that would be necessary to achieve satisfactory reliability of the mean, using the formula given by Shrout and Fleiss [8]. To check the assumption that the interval between samplings did not affect R estimates, analyses were conducted separately for measurements performed at 1 or 5 weeks interval. Similar results were observed for both intervals (data not shown). The SAS package [9] was used.

In a complementary and illustrative analysis, we used the graphical method proposed by Altman & Bland [10] which focuses on the *mean and variability of differences between pairs of repeated measurements.* (A scatter plot of the difference between the measurements against their mean allows to detect important *lack of individual reliability* which may be hidden by the use of global reliability statistics such as intraclass correlation coefficients. The plot also allows to investigate any possible relationship between the measurement error and its true value, estimated by the mean.) Plots were made, and means and standard deviations of the differences were calculated, separately for measurements performed at 1 or 5 weeks interval.

Results

Table 1 shows the distribution of studied hormonal levels. All values were included in reference intervals for normal subjects. The intraclass correlation coefficients and their 95% confidence intervals are presented in Table 2. The magnitude of the intraclass correlation coefficients for baseline levels of plasma cortisol, ACTH and LPH, nocturnal urinary cortisol and post metyrapone plasma 11-deoxycortisol ranged from 0.36–0.55, indicating only moderate reliability. Baseline salivary cortisol and plasma β-endorphin and post metyrapone levels of ACTH, LPH and plasma β-endorphin were shown to have an even lower reliability (*R*s below 0.3). Conversely, peak values of plasma cortisol, ACTH and LPH during CRH-LVP, peak values of plasma and salivary cortisol during tetracosactrin tests and post metyrapone cortisol were shown to have excellent reliability, with intraclass correlation coefficients ranging from 0.68–0.90. The use of variables transformed on a logarithmic scale did not significantly change values of *R*. In particular, intraclass correlation coefficients for baseline levels did not increase to more satisfactory levels (for example, *R* for log(plasma cortisol) and log(salivary cortisol) were 0.57 and 0.19, respectively (*vs* 0.52 and 0.18 for raw data, respectively).

Figure 1 shows the plots for pairs of measurements of baseline plasma cortisol, baseline salivary cortisol and peak value of plasma cortisol during tetracosactrin test (these selected parameters have intermediate, low and high *R* values, respectively). As expected, individual reliability was moderate or poor for baseline plasma and salivary cortisol respectively, and satisfactory (with only one exception) for peak value of plasma cortisol during tetracosactrin test. For plasma cortisol level, the standard deviation of differences was about half during tetracosactrin stimulation of that obtained at baseline. On the other hand, there did not appear to be any clear relation between the differences and the averaged hormonal levels.

The minimum number of replicate measurements that would be necessary to achieve satisfactory reliability ($R = 0.8$) of the mean is shown in Table 2. Three or four measurements would be needed for baseline levels of plasma cortisol and ACTH, urinary cortisol and the ratio 24 h urinary cortisol/creatinin and for post metyrapone plasma 11-deoxycortisol. Higher numbers of replicate measurements are necessary for the other baseline hormonal levels and for post-metyrapone ACTH and LPH levels. In contrast, a single peak value of ACTH during the CRH-LVP test, of cortisol during tetracosactrin stimulation would be sufficient. For peak values of cortisol during the CRH-LVP test, two measurements would be needed.

Discussion

We evaluated the reliability of tests commonly performed to assess the hypothalamic-pituitary-adrenocortical (HPA) activity in clinical pharmacology. Our results show that hormonal levels obtained after direct stimulation of pituitary gland (post CRH-LVP level of plasma ACTH) or adrenocortical gland (post tetracosactrin levels of plasma and salivary cortisol) have greater between-person variability, and therefore higher reliability, than those obtained at baseline evaluation. Thus a single measurement of post stimulation levels of cortisol and ACTH may suffice to characterize normal subjects in trials, although three or more baseline measurements are required to obtain satisfactory reliability of the mean for plasma cortisol, ACTH and LPH, nocturnal urinary cortisol. Despite a good inhibition of 11β hydroxylase activity (demonstrated in this study by low and reliable post metyrapone cortisol levels) post metyrapone plasma 11-deoxycortisol and ACTH also require several samplings (4 and 7, respectively) for the mean to reach a satisfactory reliability level.

Therefore baseline tests, although simpler to perform than dynamic tests, clearly appear to lack reliability in clinical pharmacology. Amongst dynamic tests, parameters obtained after direct stimulation or inhibition of the producing targets are more reliable than those obtained after indirect stimulation. These results have important implications. Low reliability increases the chance of misclassification when classi-

Table 1 Distribution of studied hormonal levels for assessing the pituitary-hormonal axis: mean, median, standard deviation (s.d.), minimum and maximum

Hormone	*Number of measurements*	*Mean*	*Median*	*s.d.*	*Minimum*	*Maximum*
*Baseline levels**						
Plasma cortisol (ng ml^{-1})†	60	146.6	137.5	41.1	76.0	247.0
Salivary cortisol (ng ml^{-1})	30	5.6	5.2	2.8	2.4	15.8
Nocturnal urinary cortisol (μg)	30	9.2	8.9	4.7	3.0	20.9
24 h urinary cortisol (μg)	30	43.1	38.0	18.1	23.0	95.0
Plasma ACTH (pg ml^{-1})†	60	35.9	31.0	23.0	11.0	98.0
Plasma LPH (pg ml^{-1})†	60	105.6	99.0	40.4	33.0	195.0
Plasma β-endorphin (pg ml^{-1})	30	27.8	24.0	13.8	12.0	75.0
Post-metyrapone levels						
Plasma 11-deoxycortisol (ng ml^{-1})	30	112.7	109.5	27.8	63.0	175.0
Plasma cortisol (ng ml^{-1})	30	48.0	41.0	39.0	2.0	166.0
Plasma ACTH (pg ml^{-1})	30	338.8	325.5	217.0	79.0	590.0
Plasma LPH (pg ml^{-1})	30	574.4	531.0	319.9	186.0	1921.0
Plasma β-endorphin (pg ml^{-1})	30	182.5	170.0	136.9	39.0	823.0
Cortrosyn test						
Peak plasma cortisol (ng ml^{-1})	30	273.2	262.5	48.9	196.0	371.0
Peak salivary cortisol (ng ml^{-1})	30	16.6	16.0	5.7	7.5	28.1
CRH-LVP test						
Peak plasma ACTH (pg ml^{-1})	30	148.1	123.5	85.6	50.0	411.0
Peak plasma LPH (pg ml^{-1})	30	246.2	241.0	101.3	111.0	531.0
Peak plasma cortisol (pg ml^{-1})	30	239.6	228.0	39.5	184.0	325.0

*Baseline levels are 08.00 h levels. †Three measurements available for 20 subjects. For other hormone levels, three measurements available for 10 subjects only. Measurements are performed at Day 1, Day 9 and Day 35 or Day 1, Day 35 and Day 44.

Table 2 Reliability of hormonal levels for assessing the pituitary-hormonal axis: intraclass correlation coefficients and minimum number of replicate measurements necessary to achieve satisfactory reliability of the mean (intraclass correlation coefficient = 0.8)

Hormone	*Intraclass correlation coefficient (95% CI)**	*Number of measurements to achieve reliability of the mean*
Baseline levels†		
Plasma cortisol‡	0.54 (0.32–0.70)	3
Salivary cortisol	0.18 (0.05–0.36)	18
Nocturnal (22.00–08.00 h) urinary cortisol	0.55 (0.18–0.80)	3
Nocturnal urinary cortisol/creatinine ratio	0.10 (0.00–0.49)	36
24 h urinary cortisol	0.28 (0.00–0.58)	10
24 h urinary cortisol/creatinine ratio	0.46 (0.14–0.71)	4
Plasma ACTH‡	0.48 (0.27–0.67)	4
Plasma LPH‡	0.36 (0.13–0.56)	7
Plasma β-endorphin	0.31 (0.00–0.60)	7
Post-metyrapone levels		
Plasma 11-deoxycortisol	0.49 (0.18–0.73)	4
Plasma cortisol	0.90 (0.80–0.96)	1
Plasma ACTH	0.33 (0.02–0.62)	7
Plasma LPH	0.26 (0.00–0.56)	7
Plasma β-endorphin	0.25 (0.00–0.56)	12
Cortrosyn test		
Peak plasma cortisol	0.84 (0.62–0.94)	1
Peak salivary cortisol	0.85 (0.64–0.94)	1
CRH-LVP test		
Peak plasma ACTH	0.90 (0.74–0.96)	1
Peak plasma LPH	0.80 (0.55–0.92)	1
Peak plasma cortisol	0.68 (0.34–0.87)	2

*CI: Confidence interval. †Baseline levels are 08.00 h levels. ‡Three measurements available for 20 subjects. For other hormone levels, three measurements available for 10 subjects only. Measurements are performed at Day 1, Day 9 and Day 35 or Day 1, Day 35 and Day 44.

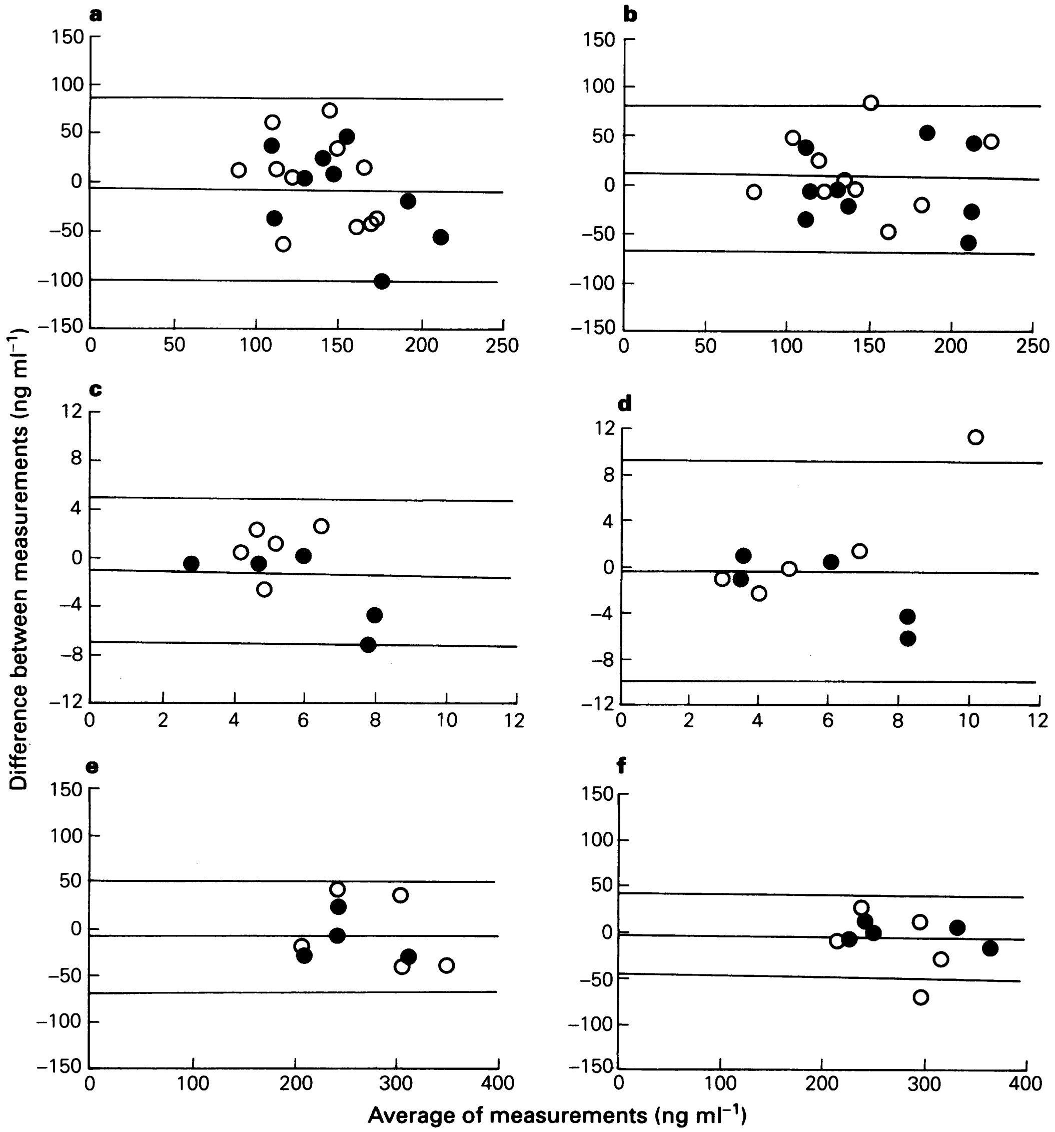

Figure 1 Difference between measurements (ng ml^{-1}) *vs* the average of measurements, n = 20 (baseline plasma cortisol) or 10 (other hormonal levels) pairs of measurements. a) baseline plasma cortisol, 1 week interval, b) baseline plasma cortisol, 5 week interval, c) basline salivary cortisol, 1 week interval, d) baseline salivary cortisol, 5 week interval e) peak plasma cortisol during tetracosactrin, 1 week interval and f) peak plasma cortisol during tetracosactrin test, 5 week interval (○ subjects with placebo in the first period, ● subjects with active treatment in the first period, the horizontal lines indicate the mean difference and mean difference ± 2 s.d.s.).

fying individuals in a study group with respect to particular thresholds or cut-off points. Moreover, unreliable measurements cannot be expected to correlate with any other variables (bias toward the null hypothesis in statistical tests), and their use in analysis often violates statistical assumptions [11].

There are at least two ways of resolving the problem of reliability. One is to choose among a series of equivalent tests those which have the lowest random-error variability (and the highest intraclass correlation coefficient). The second is, for a given test, to increase the intraclass correlation coefficient. This can be done by refining the laboratory process to decrease method variability, or by taking blood from the participant at different times and using the mean of these samples. Such multiple sampling can give an intraclass correlation coefficient as close to 1 as desired – but obviously at an increased cost. For example, the intraclass correlation coefficient for baseline plasma cortisol goes up from 0.54 with one sample to 0.81 with four independent samples.

Knowledge of the variability components of a criterion also provides help in deciding on the best design of a trial. Crossover trials have advantages compared with 'parallel group' trials when between-patient variability is high [12]. Indeed, the analysis of variance for crossover trials allows the 'elimination' of the between-patient variability, thus decreasing the residual variance proportionally to the size of the intraclass correlation coefficient [13]. In other words, the higher the intraclass correlation coefficient, the greater the increase in power of statistical analysis with the crossover design.

One limitation of our study is the small sample size, as evidenced by the magnitude of some of the confidence intervals. However, many reliability studies use relatively small samples sizes (as well as most clinical pharmacology studies). Furthermore, the importance of evaluating the functioning of the HPA system is such that any additional information on the reliability of tests is helpful, particularly when considering the complexity and costs of hormonal investigations. On the other hand, this study was conducted within an experienced clinical pharmacology unit associated with a specialized endocrine laboratory. Precautions have been taken to ensure that sera were analyzed in the same assays and by the same technician. Thus, method (analytical) errors were probably minimized. Serious errors in analyses will result if the laboratory procedures are not adequate or if the specimens have been stored or processed inappropriately.

It is required, in clinical pharmacology studies, to establish high-quality reference procedures for laboratory measurements such that data are both reliable and valid. Although validity is the usual first step toward the choice of a criterion, reliability is less commonly the subject of attention. However, it should be emphasized that poor reliability of a procedure leads to increased variance and thus poor validity estimations and misclassifications. Evaluation of the HPA system function is not an exception.

We therefore recommend that studies which include measurements of adrenocortical and corticotroph hormones should be planned and reviewed carefully with regard to the various components contributing to measurement variability.

We are greatly indebted to Drs J.L. Pinquier and P. Perles for conducting the clinical studies and Mrs C. Fournier, D. Guiban and M.F. Proeschel for their performance of hormone assays.

References

1 Streeten DHP, Anderson GH, Dalakos TG *et al.* Normal and abnormal function of the hypothalamic-pituitary-adrenocortical system in man. *Endocrine Rev* 1984; **5**: 371–294.

2 Winer BJ. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971: 248–51.

3 Kaye TB, Crapo L. The Cushing syndrome: an update on diagnostic tests. *Ann interm Med* 1990; **112**: 434–444.

4 Laudat MH, Cerdas S, Fournier C, Guiband D, Guilhaume B, Luton JP. Salivary cortisol measurement: a practical approach to assess pituitary-adrenal function. *J clin Endocrinol Metab* 1988; **66**: 34–38.

5 Laudat MH, Billaud L, Thomopoulos P, Vera O, Yllia A, Luton JP. Evening urinary free corticoids: a screening test in Cushing's syndrome and incidentally discovered adrenal tumours. *Acta Endocrinol* 1988; **119**: 459–464.

6 Kuhn JM, Bertagna X, Seurin D, Gourmelain M, Girard F. Plasma lipotropin increase in man after growth hormone administration. Comparison between extractive and biosyntethic hormones. *J clin Endocrinol Metab* 1983; **56**: 1333–1340.

7 Fleiss JL. *The design and analysis of clinical experiments.* New York: John Wiley & Sons, 1986; 8–13.

8 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; **86**: 420–428.

9 SAS Institute, Inc. *SAS/STAT users guide, version 6.* Cary, NC: SAS Institute, Inc, 1990: 893–896.

10 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 1983; **32**: 307–317.

11 Fuller WA. Measurement error models. New York: John Wiley & Sons, 1987; 7–15.

12 Pocock SJ. *Clinical trials. A practical approach.* New York: John Wiley & Sons, 1983; 110–132.

13 Wagner JG. *Fundamentals of clinical pharmacokinetics* (2nd ed.). Hamilton, Ill: Drug Intell Public, Inc, 1979; 303–305.

(*Received 7 December 1993, accepted 25 July 1994*)