

## Within-Subject Variation of Carcinoembryonic Antigen in Colorectal Cancer – Application of Reference Change Values and Individual Reference Ranges to Patient Follow-Up

Josefina Mora and Josep M. Queraltó

Servei de Bioquímica, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

**Summary:** Dynamic interpretation of results is an alternative approach to the conventional cut-off procedure. Reference change limit is a valuable reference point to interpret dynamically tumour marker values when only very few serial results can be obtained from a patient after treatment. In this paper, a reference change limit of 2.0 µg/l was estimated for carcinoembryonic antigen in patients with complete remission of colorectal cancer. This figure means that a difference greater than 2.0 µg/l has at least a chance of being statistically significant (at 0.05 probability). As complementary information to the reference change limits, with more than four successive results, a simple time series model can be used to obtain predictive limits for the next observation.

### Introduction

When monitoring cancer patients for possible recurrence after treatment, it would be helpful to have reference points derived from the individual patient or, at least, from a group of patients in remission. A reference point of the latter type is the “reference change limit” or “critical difference” to judge the statistical significance of a change between two successive observations of a tumour marker. The reference change limit is particularly valuable when only a very few serial values have been obtained after treatment. When the patient’s remission period has extended long enough to provide a larger number of observations, predictive limits derived from a time series model applied to the patient’s own record represent the best reference points-patient-specific values (1).

There is growing interest in within-subject variability. This information is important in calculating reference change limits, assessing the appropriateness of conventional reference values, calculating intra-individual reference values, establishing analytical goals, and helping in test and specimen selection (2). Most of these studies have been done in groups of healthy subjects (2–4) although there are some published data on biological variation in stable patients with chronic disease (5). Interpretation and, consequently, usefulness of tumour marker results remains controversial (6). At present, reference points for tumour markers are usually the conventional group-based cut-off points derived from samples of healthy persons. Several authors have claimed that a more dynamic approach would be better (7–8). If criteria were based on relative increase, the interpretation would improve dramatically (7–18). Therefore, it

would be possible to consider irrelevant increases of differences in concentration under the reference change limit, or suspect early recurrence when an increase in concentration (not due to biological variation) is detected.

In this paper, we estimate a group-based reference change limit and patient-specific predictive limits for carcinoembryonic antigen in patients in remission following treatment for colorectal carcinoma. These reference points are then tested against the CEA records of patients in clinically evident pre-progressive and progressive stages of the disease, as a preliminary approach.

### Materials and Methods

#### Patients

Patients in complete remission from colorectal cancer following treatment were selected among patients attending the Hospital de la Santa Creu i Sant Pau, Barcelona, Spain, during the period January 1990 to December 1992. Selection criteria included:

- (a) clinical criteria: minimum of three months since last treatment (history, examination, every 4 months);
- (b) complementary explorations (colonoscopy at 4 and 36 months; ultrasound and chest X-ray film at 1, 2 and 3 years) (19);
- (c) at least four regularly-spaced carcinoembryonic antigen results (averaging three months apart, but see tab. 1) following treatment while the patient remained in steady remission.

In order to have a homogeneous group where the only factor under study was time, patients with benign pathology or presenting factors able to modify the carcinoembryonic antigen concentration were excluded. The group numbered 30 patients (20 men and 10 women, aged between 41 and 78 years, stages A (3), B (15), C (14), D (1), according to the *Astler-Coller* modified classification and histology of adenocarcinoma), accounting for 117 successive observations. At approximately five years after surgery, the patient records were revised and all remained in remission.

We also selected 17 patients [19 men and 7 women, aged 26–78 years, stages B (4), C (7), D (6)] with colorectal adenocarcinoma

confirmed clinically or by image exploration (19) after surgical treatment but in progression or recurrence (20) in order to test how several calculation approaches to reference changes can modify the conclusions.

#### Collection of blood samples and analytical methods

The mean interval between sample collections for each patient was 3.39 months with a range from one to seven months. However, actual intervals varied considerably, as shown in tables 1 and 2.

Venous blood was obtained by antecubital puncture from seated subjects and collected directly into plain tubes (10 ml) between 8–10 a.m. with minimal stasis. Serum specimens were obtained by centrifugation at 10 000 g for 10 minute within one hour after arrival in the laboratory. Samples were stored at  $-20^{\circ}\text{C}$  until analyzed, not later than 30 days, and carcinoembryonic antigen concentration was determined by an enzyme-chemiluminometric method (Amerlite<sup>®</sup>, Kodak Clinical Diagnostic Ltd., Amersham International, Buckinghamshire, UK). This assay involves the simultaneous reaction of carcinoembryonic antigen with sheep polyclonal antibodies against carcinoembryonic antigen-coated wells and peroxidase-labelled mouse monoclonal antibodies against carcinoembryonic antigen. Peroxidase activity is measured by

**Tab. 1** Relative frequency distribution (in percent) of the number of blood specimens obtained (columns labelled 1 to 8) in relation to the number of months elapsed between consecutive observations (rows labelled 1 to 7).

Interval from the previous observation	Number of observations							
	1	2	3	4	5	6	7	8
1	4.0	20.0	12.0	0.0	0.0	8.3	0.0	0.0
2	48.0	32.0	24.0	25.0	13.3	16.7	0.0	0.0
3	24.0	8.0	28.0	15.0	33.3	25.0	25.0	0.0
4	12.0	24.0	12.0	20.0	40.0	25.0	50.0	100
5	8.0	8.0	12.0	25.0	13.3	16.7	25.0	0.0
6	4.0	8.0	12.0	5.0	0.0	8.3	0.0	0.0
7	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0
Total number of patients	25	25	25	20	15	12	4	1

**Tab. 2** Relative frequency distribution of months elapsed between blood specimen collection of all the successive differences observed, and frequency distribution of the average of time elapsed (month) per patient.

Interval (in months) from the previous observation	Frequency distribution of differences observed	
	All differences observed	Mean of elapsed time (months) per patient
1	7.9	0.0
2	27.6	4.2
3	21.3	45.8
4	22.0	29.2
5	13.4	12.5
6	5.5	8.3
7	2.4	4.2
Total number of differences	127	25
Mean (standard deviation)		3.39 (1.14)

enhanced luminescence reaction using a luminol derivative as substrate.

Three control specimens from Lyphocheck<sup>®</sup> (Bio-Rad, ECS Division, Anaheim, California 92806) and one human serum pool aliquot were included in each assay run. Different lots of control material were used throughout the study period. Coefficients of variation of the control material assays during this period were: 9.86, 7.17 and 8.55 percent for 3.24, 15.18 and 31.73  $\mu\text{g/l}$  respectively ( $n = 114$ ). Assay procedures remained unchanged throughout the study period.

#### Statistical analysis

The primary purpose of this analysis was to determine critical values (reference change limits) for use in evaluating observed differences between successive observations early in the monitoring process. In addition, we calculated patient-specific predictive limits after at least four serial observations had been obtained. Because of the wide variation in sampling intervals and the relatively small number of patients, the only time series model available to us was the simple "homeostatic" model (1). This model assumes zero (or very small) correlation between successive results. A much larger database ( $> 300$  patients) would have permitted the use of a continuous autoregressive model allowing for unequal intervals and stronger serial correlation (21).

These calculations are based on the premise that the patients remained in complete remission during the monitoring period. Under this hypothesis, both reference change limits and predictive limits become reference points which, if exceeded, indicate that the patient may have experienced a recurrence of the disease. In practice of course, patients may show random deviations beyond these reference points that prove on further monitoring or other examination to be false alarms. However, five patients showed consistent upward trends during their first four observations (for example: 0.50, 0.70, 2.10, 2.50  $\mu\text{g/l}$ ) although none of them ever showed values over 5  $\mu\text{g/l}$  and some time after they recovered previous values. These were not considered in calculating the reference change limit. This reduced the database to 25 patients.

#### Calculation of Reference Change Limits

We turn now to the determination of a reference change limit suitable for judging the statistical significance of a difference between two successive observations. This will be useful during early monitoring when no more than 2–4 serial observations have been obtained. The reference change limit is calculated from the data for all patients in the group, whereas predictive limits are calculated for each patient individually.

Calculation of reference change limits was initially proposed by *Harris & Brown* (4). Since that time, reference change limits have been computed for many different analytes, not always using the procedure originally suggested. A discussion of this experience may be found in *Harris & Boyd* (22). In brief, this technique involves:

- checking the distribution of observed within-person variances for assumed log-normality;
- estimating the mean and standard deviation of these variances (on the log scale) using *Healy's* trimming procedure (23) to obtain robust estimates in the presence of outliers;
- converting these estimates to original scale using standard formulas [e.g. 1.c. (24) and 1.c. (4)] testing

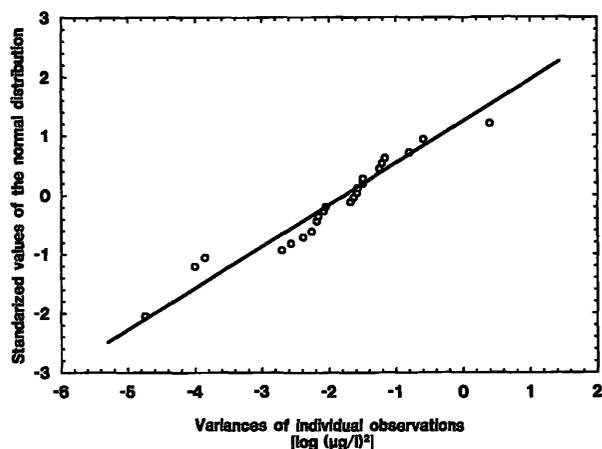


Fig. 1 Distribution of log-variances in a probit scale.

whether observed variances are homogeneous (i. e., estimates of a common true variance). If not, the standard deviation of true variances is estimated [(24) equation 3]. From this estimate and the estimated mean of true within-patient variances, the reference change limit is calculated. Further details and a BASIC program for these computations are given in *Harris & Boyd* (22).

First, we examined the distribution of within-person variances in the 25 patients referred to above. Variances were computed from all observations of each patient because as far as is known all patients remained in complete remission during the study period. Two of the 25 patients showed unstable records with unusually large variances. These were omitted from the reference change limit calculations. The distribution of the remaining 23 variances seemed to be in reasonable agreement with a logNormal form as shown in figure 1, with the possible exception of the lowest and highest values. These were (temporarily) deleted for application of *Healy's* procedure, leaving 21 covariances, based on an average of 6.0 observations per patient.

The true within-person variances were found to be non-constant with estimated CV of 99%. Based on the estimated 90th percentile of the distribution of true variances, a reference change limit of 1.97 µg/l was obtained and rounded off to 2.00 for convenience<sup>1)</sup>. The inter-

<sup>1)</sup> The BASIC program used here does not allow for varying numbers of serial observations per patient. Therefore, as a check on the reference change limit, we computed the weighted mean and variance of observed within-patient variances (weighting by the degrees of freedom for each patient), omitted *Healy's* procedure, and estimated the variance of true variances using the following formula [l. c. (22) p. 196 and l. c. (25)] that accounts for each patient's number of observations:

$$\text{est Var } \sigma_i^2 = [\text{var } s_i^2 - 2 \times A \times (\text{mean } s_i^2)^2]$$

where A is the average value of  $1/(n_i - 1)$ . Then, following the remaining portion of the program, the calculated reference change limit came to 1.91 µg/l, in approximate agreement with the value 1.97 computed above. It appears that minor variation in the number of observations per patient can be safely ignored, substituting the mean number instead.

pretation of this value is that any difference greater than 2.0 has at least a 90 percent change of being statistically significant at the 0.05 probability level. Applying this reference change limit to the 114 successive differences among the observations in the 23 patients, we found only one difference (from 5.24 to 2.31 µg/l) that exceeded this value (which means 0.9 percent of false alarms or a specificity of 99.1 percent of differences studied, or 95.6 percent in terms of patients). Of the 12 differences in the two patients with unusually unstable records, four exceeded this reference change limit. Of the 50 differences observed in 17 patients with disease progression, only two differences were under 2 µg/l (a sensitivity of 96 percent of differences of 11.8 percent in terms of patients). In this group, four patients showed the first observations under the conventional static 5 µg/l cut-off and two of them showed again a result under 5 µg/l in the second test. However, the difference between the first and second observations was higher than the reference change limit in these two cases. Consequently, progression could be suspected 8 and 3 months respectively before a new assay (the results of which were respectively 4.7 and 6.5 µg/l) was performed. If, instead of applying the 90th percentile as the calculation procedure, we had simply assumed the 23 observed within-patient variances to be homogeneous and used their mean (0.2909) as the best estimate of the true variance, we would have calculated the reference change limit as  $2.77 (0.2909)^{1/2}$ , or 1.49 µg/l. This value was exceeded by 5 of the 114 differences observed among the 23 individuals (that means 4.4 percent of false alarms or a specificity of 95.6 percent). There were no differences in the group of patients with disease progression. Finally, if, as is commonly done, the median (0.2274) had been used, the calculated reference change limit would have been 1.32 µg/l, a value exceeded by 6 of these differences (that means 5.3 percent of false alarms or a specificity of 94.7 percent). Although the population of patients with disease progression is limited to have a good sensitivity estimation, the better specificity is obtained with a reference change limit calculated as proposed originally (22).

The five patients provisionally excluded from the calculations showed a mean of an intra-individual variances of 0.40. The highest of the 27 differences observed in the patients of this group was 1.40 (from 0.70 to 2.10 µg/l). The inclusion of this group would slightly modify the reference change limit from 1.97 to 2.07 µg/l.

### Individual Predictive Limits

After four successive results at approximately equal time intervals have been obtained, the reference change limit may be replaced, or at least supplemented, by a simple time series model used sequentially to obtain predictive

limits for future observations in the individual patient. Details of this methodology, and relevant references, appear in *Harris & Boyd (22)*. Briefly, given four serial observations, not necessarily equally spaced, while the patient remains in steady state, and assuming zero (or very small) serial correlation, the best estimate for the fifth observation is the mean of the first four, with 95 percent predictive range given by this mean  $\pm$  "t"<sub>0.025, 2</sub> s<sub>4</sub> (5/4)<sup>1/2</sup>, where "t"<sub>0.025, 2</sub> denotes the two-sided value of *Student's t* at two degrees of freedom and the 0.05 probability level. In general, provided the patient remains in steady state, and the serial correlation of successive observations remains very low, the best predictor of the t-th observation, based on (t-1) observed values, and associated 95 percent predictive range is given by:

$$x_{t-1} \pm "t"_{0.025, (t-3)} \times s_{t-1} \times \sqrt{\frac{t}{t-1}} \quad (\text{Eq. 1})$$

Where  $x_{t-1}$  and  $s_{t-1}$  are the mean and standard deviation, respectively, of the first (t-1) observations. Since only higher than usual observations are of clinical importance in this application, only the upper predictive limit should be computed and a one-sided value of "t" used to increase sensitivity.

Including all patients with at least five serial observations, but excluding those showing a significant trend in the first four observations, 20 patients remained. For each of these patients, equation 1 was used to calculate upper predictive limits for the fifth and later observations. Of the 56 observations tested, only 4 exceeded their respective upper predictive limits: 1.19  $\mu\text{g/l}$  for a predictive interval of (0.35–0.68  $\mu\text{g/l}$ ); 1.97 (1.0–1.79); 1.09 (0.12–0.90) and 1.8 (0.1–1.25). Therefore, the rate of false positive signals is 7 percent. None of the 5 cases excluded from the reference limit change calculation exceeded their respective upper predictive limit.

## References

- Harris EK. Some theory of reference values. II Comparison of some statistical models of intraindividual variation in blood constituents. *Clin Chem* 1976; 22:1343–50.
- Fraser CG, Harris EK. Generation and application of data on biological variation in clinical chemistry [review]. *Crit Rev Clin Lab Sci* 1989; 27:409–37.
- Boyd JC, Harris EK. Utility of reference change values for the monitoring of inpatient laboratory data. In: Zinder O, editor. *Optimal use of the clinical laboratory*. 5th Int. Meet. Clin. Lab. Organization and Management, Haifa August 26–29, 1985. Basel: Karger, 1986:111–22.
- Harris EK, Brown SS. Temporal changes in the concentration of serum constituents in healthy men. Distributions of within person variances and their relevance to the interpretation of differences between successive measurements. *Ann Clin Biochem* 1979; 16:169–76.
- Hölzel WGE. Intra-individual variation of some analytes in serum of patients with chronic liver disease. *Clin Chem* 1987; 33:1133–6.
- Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen C. An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected colon cancer. *J Am Med Assoc* 1993; 270:943–7.
- Gion M; Ruggeri G, Marconato R, Rasella C, Nosadini A, Lafranchini G, et al. Perioperative kinetic evaluation of tumour markers in breast and colorectal carcinoma. *J Nucl Med Allied Sci* 1990; 34 Suppl 4:9–12.
- Kiang DT, Greenberg LJ, Kennedy BJ. Tumour marker kinetics in the monitoring of breast cancer. *Cancer* 1990; 65:193–9.
- Minton JP, Martin EW. The use of serial CEA determinations to predict recurrence of colon cancer and when to do a second-look operation. *Cancer* 1978; 42:1422–7.
- Winkel P, Lundbye-Christensen S, Christensen M. Analysis of serial measurements in clinical practice. *Scand J Clin Lab Invest* 1992; 52 Suppl 208:105–21.

The expected number of such false alarms is given by simple binomial theory. The predictive value for the observation  $x_t$ , conditional on all preceding observations being in steady state, is called a one-step-ahead forecast. All one-step-ahead forecasts are mutually uncorrelated (25), and, in the case of normally distributed observations, independent. Therefore, each successive test of the deviation between  $x_t$  and its predictive limit represents an independent binomial trial. The expected number of false alarm probability for each, assuming continuation of the steady state, is then the total number of tests times the false alarm probability for each, in this case 56 times 0.05, or 2.8. Thus, our results appear to conform reasonably well to expectations.

In a preliminary evaluation, as expected, most of the successive values from the 17 patients in progression showed increments higher than the reference change limit. However three patients (accounting for eleven observations, seven differences) showed five significant differences in values near the cut-off (5  $\mu\text{g/l}$ ), several months before clear departures of the steady state. Furthermore, all patients showing values below the cut-off before recurrence also showed differences between successive observations below the reference change limit. However, in these patients there was a clear trend preceding the time of evident progression.

In conclusion, as tumour markers generally show a high degree of individuality, cut-offs derived from the healthy population may be of limited value in assessing unusual results in a patient follow-up (27). Therefore, the use of reference change values and predictive individual limits may be of valuable help in assessing any increase in tumour marker concentration during follow-up, providing there is no trend in the data.

## Acknowledgements

We thank *Eugene K. Harris* for his valuable suggestions, helpful discussions and for the critical review of the manuscript.

11. Ravry M, Moertel CG, Schutt AJ, Go VL. Usefulness of serial serum carcinoembryonic antigen (CEA) determinations during anticancer therapy or long-term follow-up of gastrointestinal carcinoma. *Cancer* 1974; 34:1230-4.
12. Riesen W, Kessler AC, Ehrhardt V. Individual variations of carcinoembryonic antigen (CEA) in serum of healthy subjects. *Clin Chem* 1987; 33:2123.
13. Browning MCK. Biological variation of the breast tumour marker CA 15.3: implications for necessary standards of performance and the interpretation of results. *Ann Clin Biochem* 1987; 24 Suppl. 1:35-8.
14. Browning MCK, McFarlane MP, Morobin JM, Preece PE, Wood RAB. Evaluation of the comparative clinical utility of CA 15.3 and mucinous carcinoma-associated antigen (MCA) in the management of breast carcinoma. *Ann Clin Biochem* 1988; 25:54-7.
15. Panteghini M, Pagani F, Bonora R. Pre-analytical and biological variability of prostatic acid phosphatase and prostate-specific antigen in serum from patients with prostatic pathology. *Eur J Clin Chem Clin Biochem* 1992; 30:135-9.
16. Gion M, Cappelli G, Mione R, Vignati G, Fortunato A, Saracchini S, et al. Variability of tumour markers in the follow-up in patients radically resected for breast cancer. *Tumour Biol* 1993; 14:325-33.
17. Söletormos G, Schioler V, Nielsen D, Skovsgaard T, Dombernowsky P. Interpretation of results for tumour marker on the basis of analytical imprecision and biological variation. *Clin Chem* 1993; 39:2077-83.
18. Hölzel WGE, Beer R, Deschner W, Griesmacher A, Müller MM. Individual reference ranges of CA 15-3, MCA and CEA in recurrence of breast cancer. *Scand J Clin Lab Invest* 1995; 55 Suppl 221:93-101.
19. Cohen AM, Minsky BD, Schilsky RL. Colon cancer. In: DeVita VT, Hellman S, Rosenberg SA, editors. *Cancer: principles and practice of oncology*. 4th ed. Philadelphia: J. B. Lippincott, 1993:929-77.
20. Müller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981; 47:207-14.
21. Schlein BR, Lavin PT, Hayden CL. Using an autoregressive model to detect departures from steady states in unequally spaced tumour biomarkers data. *Stat Med* 1992; 11:515-32.
22. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York, NY: Marcel Dekker Inc., 1995:221-58.
23. Healy MJR. Outliers in clinical chemistry quality control schemes. *Clin Chem* 1979; 25:675-7.
24. Queraltó JM, Boyd JC, Harris EK. On the calculation of reference change values, with examples from a long-term study. *Clin Chem* 1993; 39:1398-403.
25. Harris EK. Distinguishing physiologic variation from analytic variation. *J Chron Dis* 1970; 469-80.
26. Box GEP, Jenkins GM. Time series analysis, forecasting and control. San Francisco: Holden Day, 1970:129.
27. Deijter SW, Martin JS, McPherson RA, Lynch JH. Daily variability in human serum prostate-specific antigen and prostatic acid phosphatase: a comparative evaluation. *Urology* 1988; 32:288-92.

*Received September 9, 1996/April 14, 1997*

Corresponding author: Josep M. Queraltó, Servei de Bioquímica, Hospital de la Santa Creu i Sant Pau, Av. Sant Antoni M. Claret, 167, E-08025 Barcelona, Spain

