Primary unit for statistical analysis in morphometry: patient or cell?

A methodological investigation on the basis of 16 oxyphilic follicular neoplasms of the thyroid

Oleksiy Tsybrovskyy [a,*] and Andrea Berghold [b]

[a] *Department of Pathology, University of Graz, School of Medicine, Graz, Austria*
[b] *Department of Medical Informatics, Statistics and Documentation, University of Graz, School of Medicine, Graz, Austria*

Received 3 December 1998
Accepted 2 July 1999

In a series of 16 oxyphilic follicular neoplasms of the thyroid (8 adenomas and 8 carcinomas), three different approaches for the analysis of morphometric data were evaluated. It was shown that the statistical design of morphometric studies is by nature nested due to subsampling of cells within each patient. Therefore, the most appropriate analysis would be to account for this hierarchical structure. However, related statistical methods are not at present well established, especially as far as classification rules are concerned. Therefore, the nested design is converted into the simple factorial one by considering only one kind of statistical unit – either patients or cells. The results of the study presented indicate that ignoring the patient as unit of analysis leads to a substantial error in statistical output, regardless of the particular procedure applied. Moreover, the size of the error can be neither diminished nor controlled. Choosing patients as primary units assures accurate results and also has an advantage of gaining some additional information by calculating several distributional estimates in each patient. However, this approach often requires a reduction of dimensions and, furthermore, is not encouraged in certain fields of quantitative cytology. Advantages and disadvantages of all approaches have been summarized and practical recommendations for their use have been worked out.

Keywords: Computer assisted image analysis, statistical analysis, nested design, thyroid, oxyphilic neoplasms

*Correspondence to: Oleksiy Tsybrovskyy, c/o Helmut Denk, M.D., FRC Path., Department of Pathology, University of Graz, Auenbruggerplatz 25, 8036 Graz, Austria. Tel.: +43 316 380 44 01; Fax: +43 316 38 43 29.

Analytical Cellular Pathology 18 (1999) 191–202
ISSN 0921-8912 / $8.00

1. Introduction

Recent advances in computer science have led to a substantial progress in morphometry [27,43]. However, numerous methodological problems still remain, statistical issues among them [27,43]. We came across many of these problems while performing our recent work on thyroid neoplasms [44]. In our hands, one of the most challenging problems deserving special attention was an appropriate choice of the primary unit for statistical analysis.

The problem starts with the fact that usually two different sampling units exist in morphometry: patients (tumours) and cells (nuclei).[1] With regard to the statistical evaluation, three different methods of analysis ensue, which all can be illustrated by corresponding examples from the literature. The first approach (approach I) takes into account *both* units, i.e., a so-called "nested" design is considered [6,19], where patients represent "higher-level" units and cells "lower-level" units (Fig. 1A). Some special techniques, such as mixed-effect analysis of variance (ANOVA), have been adopted for this design [6,8,13,24].

The second approach (approach II), which is widely used in combination with discriminant analysis, is to treat separate *cells* as primary units [8,15,22,24,34,35, 40], thereby ignoring patients as units of analysis. That is, nuclei from all tumours are pooled, for instance, into "malignant" and "benign" groups (Fig. 1B), and then, a multivariate "cell classifier" is developed. As a result, each individual nucleus is assigned by the program to a particular diagnostic group. However, to achieve practically meaningful results a subsequent classification of tumours has to be made [12]. This is usually based either on the mere prevalence (50%) of cells assigned to a given diagnostic group [22] or on the use of more

[1] Here and further we use the words "cell" and "nucleus" as well as "tumour" and "patient" interchangeably, because in the context of the present paper their meanings are identical.

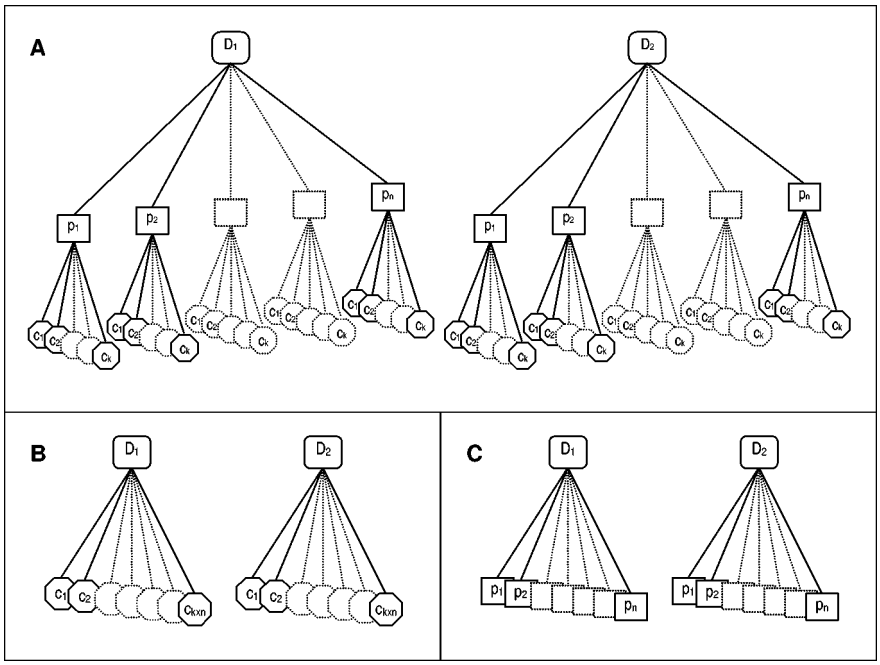

Fig. 1. Different kinds of statistical design in morphometry: A, nested design with one fixed (diagnosis) and one random (patients) factor; B, simple factorial design with cells as analytical units; C, simple factorial design with patients as analytical units. D_1, D_2, diagnostic groups, p, patients, c, cells, n, number of patients in each diagnostic group, k, number of cells measured in each patient.

flexible cut-off values [12,34,35]. Thus, in approach II a full run of multivariate analysis is carried out exclusively on separate cells and only thereafter a tumour is regarded as a whole.

In the third approach (approach III) *patients* are considered as primary units (Fig. 1C). Commonly, average values and standard deviations of nuclear features are determined in every tumour [3,21,25,37,38,41,44,45]. A number of other distributional indices (such as skewness, kurtosis, Gaussian and Fourier components) can be calculated as well [5]. Values obtained comprise the "final" data set, on which all further statistical analysis, including the development of classification models, is carried out.

The correct choice among these approaches is difficult. To our surprise, we failed to find any discussion of this problem in the literature while doing the previous work on non-oxyphilic thyroid neoplasms [44]. We therefore decided to carry out a methodological investigation on an independent exemplary data set derived from 16 oxyphilic thyroid tumours. This is also the reason why only papers devoted to surgical pathology of the thyroid are cited above for presentation of the approaches. Nevertheless, the same situation can readily be observed in any other field of diagnostic morphometry, and the problem is thus of general character. Hence, the major aim of our study was to compare the described ways of analysis and to work out a "safe choice" strategy, which can help to avoid some erroneous conclusions in future morphometric investigations.

2. Materials and methods

2.1. Case selection and measurements

Sixteen oxyphilic follicular neoplasms of the thyroid (8 adenomas and 8 carcinomas, archival material) were analyzed. The diagnoses were based on WHO criteria [29]. All specimens were fixed in buffered 10% formalin and embedded in paraplast. In every case, new 5-μm thick Feulgen stained sections were made from a block with representative tumour areas. Measurements were performed by means of a semi-automatic system for image analysis (CIRES, KONTRON, Germany), as described previously [44]. In each tumour, 200 to 350 randomly selected nuclei were measured using the following seven parameters [14]: nucleus area (NA); mean optical density within a nucleus (MOD); standard deviation of optical density within a nucleus (SDOD); skewness in the frequency distribution of op-

tical density within a nucleus (SkewOD); excess in the frequency distribution of optical density within a nucleus (ExcOD); minimal optical density within a nucleus (MinOD); maximal optical density within a nucleus (MaxOD).

2.2. Statistical analysis

Statistical analysis was performed using SPSS for Windows, version 6.1.3a (SPSS Inc., USA). To simplify the calculations, we created a completely balanced design, i.e., we left for the analysis only 200 nuclei per patient by a random elimination of all surplus cells. Two methods, analysis of variance (ANOVA) and linear discriminant analysis (DA), were used to evaluate possible differences between the tumour groups. The underlying assumptions (normality and equality of covariance matrices) were always checked graphically, both before analysis (on frequency histograms) and after (on residual plots) [1,2,18,39]. The three approaches, which we will continue to call "approach I", "approach II" and "approach III", were explored independently from each other as follows.

Approach I. To account for the nested structure of the data, a two-way ANOVA with mixed effects was applied. In this model (model 1), patients represented the first, random effect (i.e., a factor with infinite number of levels), which was nested within the second, fixed factor, i.e., diagnosis [6,13] (Fig. 1A). As for DA, no well-established adaptations of this technique to nested design seem to exist, so this kind of analysis was not attempted.

Approach II. After pooling cells within each diagnostic group, there were $k \times n = 1600$ observations in the "benign" and 1600 in the "malignant" group (Fig. 1B). For hypothesis testing, a one-way ANOVA (model 2) was performed. In DA, stepwise selection of variables, based on minimization of Wilks' lambda (F for entry 3.84, F for removal 2.71), was adopted [18, 39]. The fit of the final discriminant function was estimated on the same data set using the "jack-knife" (leave-one-out) method. That is, nuclei belonging to a particular patient were excluded from the analysis and then classified by the discriminant function computed from remaining observations, with repetition of the procedure for each patient in turn [18,39]. The subsequent tumour classification as well as probabilities of the predicted diagnoses were based on the prevalent percentage of cells assigned by the model to a given category.

Approach III. In this approach (Fig. 1C), values of all parameters were averaged separately for each patient. For comparison with the first two approaches, we subjected the mean values to one-way ANOVA tests (model 3). To extract more information from the data, several other indices (standard deviation, skew, kurtosis) characterizing the shape of frequency distribution for every feature within a given tumour were also calculated [5]. Thus, the final data set contained four times as many variables (28) as the original one. However, we could not perform multivariate ANOVA on such a data set directly, because there were more variables than observations and the covariance matrices were therefore redundant. For DA, the ratio observations vs. variables (ROV) was even more critical. Indeed, for discriminant function to be accurate enough, the ROV value must be at least 2 (optimally 5–10 or more) [4,18]. It must be stressed that the calculation of ROV involves not the general number of observations, but only the number in the smallest group [4,18]. Hence, the initial ROV in our study ($8 : 28$) was unacceptable. To overcome this problem, we performed a reduction of dimensions not involving the outcome variable [28,36]. At first, a hierarchical clustering of variables using Ward's method was performed. As a dissimilarity measure, the squared Euclidean distance was specified. To equalize effects of differently scaled variables, all values in the data set were standardized to a range of 0 to 1 [39]. As shown in Fig. 2, all variables fell into three clusters of about equal size. Each separate cluster was then subjected to factor analysis. The number of extracted principal components was determined on scree plots (Fig. 3). To save maximum information, the beginning of the "scree" was always chosen as the last downright "knee" of the plot, at eigenvalues lower than 1 (arrows in Fig. 3); afterwards, all factors lying before this point were extracted. For the same reason, a non-orthogonal rotation of the factor matrix (Oblimin, delta $= 0$) was adopted [39]. As a result, 10 new variables, which accounted for an overwhelming part of the initial variance (94.7, 99.4 and 89.9% in the first, second and third clusters, respectively), were calculated. The multivariate ANOVA was already possible (model 4), but in DA, a further dimensionality reduction had to be performed, because the ROV value was still too small ($8 : 10$). This was achieved by a stepwise selection of variables (for settings, see approach I). With the two selected variables (fac-2.3 and fac-3.1), the ROV was quite reasonable ($8 : 2$). To validate the model, the "jack-knife" method (with subsequent exclusion of every patient) was used.

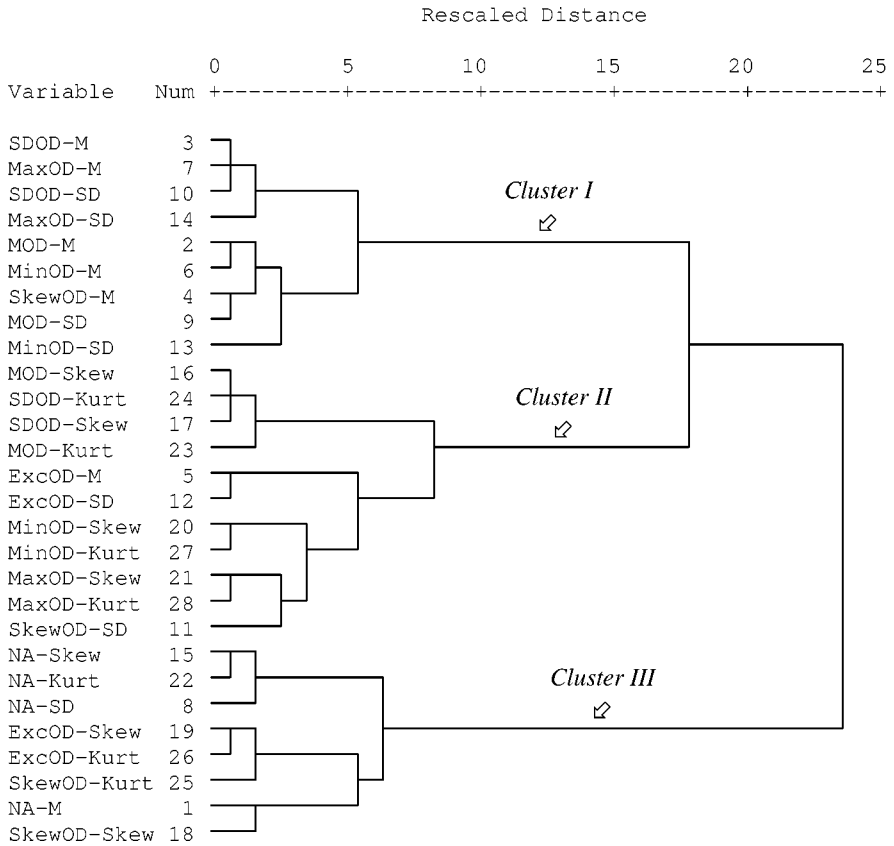

Fig. 2. Horizontal dendrogram representing steps in hierarchical clustering of karyometric parameters. The actual distances at which separate clusters were joined are proportionally rescaled to a range from 0 to 25 so that the ratio of the distances between steps is preserved. Obviously, all variables fall into three different clusters of about equal size.

3. Results

3.1. ANOVA

Results of ANOVA tests are given in Table 1. As can be seen, the first two models yielded quite different results. Indeed, model 2 demonstrated highly significant differences between "benign" and "malignant" cell populations in both multi- and univariate tests, but model 1 indicates that most of the differences can be explained by patient-to-patient variability (all tests for the patient level were highly significant), so that the overall contribution of the second factor – diagnosis – was non-significant. The controversy was especially striking in the multivariate tests. On the contrary, models 1 and 3 appeared almost identical with regard to the diagnosis level, since they produced practically equal F- and p-values. As for model 4, it incorporated a good deal of extra information brought in with the additional indices calculated. This information turned out to be important for separation of the tumour groups, because the results of multivariate tests in model 4 were statistically significant, in contrast to those in model three.

3.2. Discriminant analysis

The primary output of DA is presented in Table 2. Both discriminant functions computed were highly significant. However, the Wilks' lambda in approach II was noticeably large, suggesting a low practical efficiency of the corresponding discriminant function [4,18,39]. In reality, this flaw expressed itself through rather inconclusive probabilities for correct diagnoses when classifying individual tumours in the "jack-knife" procedure (Fig. 4). Indeed, most probability values were between 0.6 and 0.7, which is fairly low. On the contrary, the discriminant function in approach III had a relatively small Wilks' lambda (Table 2) and classified the patients rather unequivo-

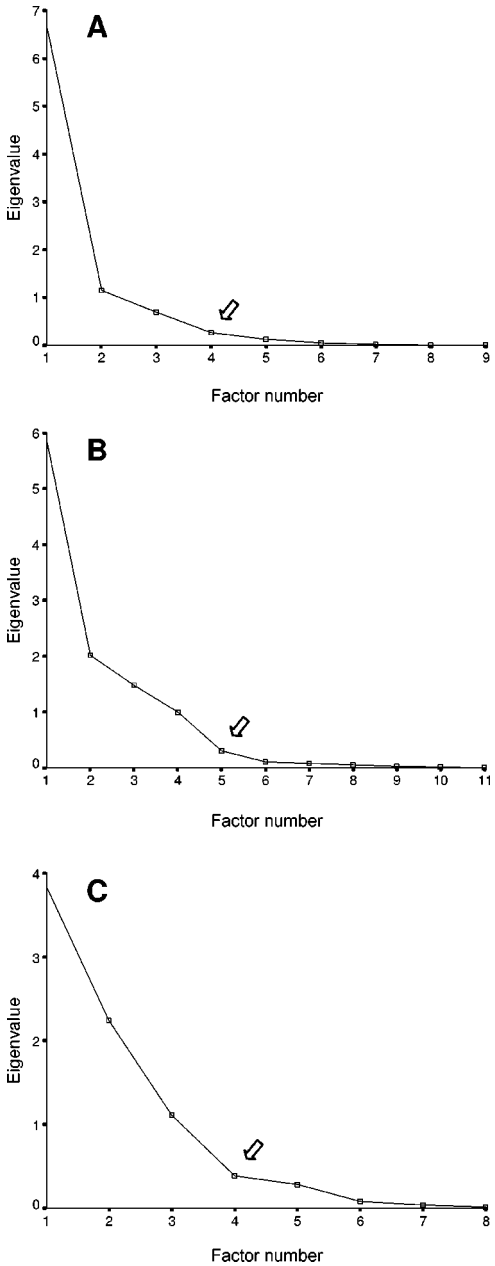

Fig. 3. Scree plots obtained in factor analysis procedure: A, in the first cluster; B, in the second cluster; C, in the third cluster. Arrows denote the beginning of the "scree".

cally (Fig. 4). A comparative performance of the classification models in terms of misclassification rates and some related indices (i.e., sensitivity and specificity for malignancy detection) after the "jack-knife" validation are given in Table 3. Here, the advantage of the approach III is not as striking, but still evident.

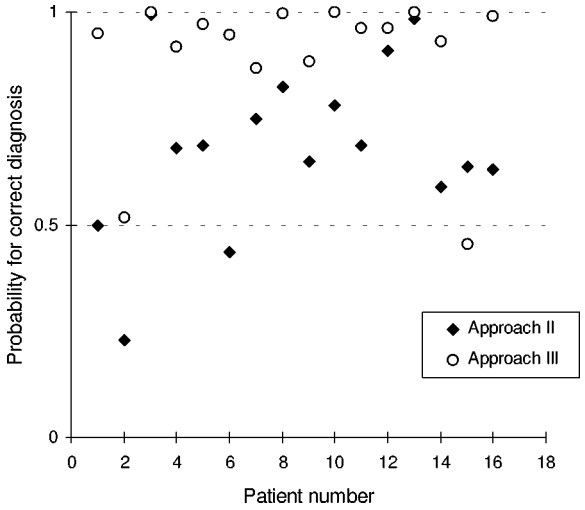

Fig. 4. Probabilities for correct diagnoses obtained in approach II and approach III after the "jackknife" validation. Points lying under 0.5 probability level correspond to falsely classified tumours.

4. Discussion

As mentioned in the introduction, there are two main sampling units in morphometry, patients and cells, which together build a hierarchically organized structure[2] (Fig. 1A). This structure is usually called "nested design" [6] and occurs as a result of the corresponding sampling procedure: first, one carries out a random sampling of patients in each diagnostic group, and then a subsampling of cells within each tumour. It should be noted that the subsampling of observations within the same objects is not rare in medicine. For instance, in periodontal research scientists also have to deal with hierarchically organized data, since many teeth (sites) per patient are usually measured. Unlike morphometry, the statistical issues in periodontology have been discussed in detail [10,16,17,20,32,33], and we shall widely use results alluded to in the discussion in our paper.

It is striking, that different ways of analyzing essentially *the same data set* produce quite different results. We shall therefore discuss the underlying causes for this discrepancy, considering each of the approaches used.

[2]For simplicity reasons, we will further discuss only a completely balanced design, i.e., when all diagnostic groups contain an equal number of patients and the number of cells measured in each patient is the same.

Table 1
Results of ANOVA-tests in all approaches

Models and effects	Tests	Variables	Hypoth. SS	Error SS	Hypoth. mean SS	Error mean SS	F	p
Model 1 (nested design, Fig. 1A)								
patient level	Univariate: $df_h = 14$; $df_e = 3184$	NA	214488.7	584926.2	15320.6	183.71	83.40	<0.0001
		MOD	9.59	16.99	0.68	0.0053	128.32	<0.0001
		SDOD	1.69	2.20	0.12	0.0007	174.86	<0.0001
		SkewOD	145.75	554.05	10.41	0.17	59.83	<0.0001
		ExcOD	103.78	1095.26	7.41	0.34	21.55	<0.0001
		MinOD	1.15	5.51	0.082	0.0017	47.53	<0.0001
		MaxOD	51.15	85.87	3.65	0.027	135.49	<0.0001
	Multivariate:							
	Pillais						47.35	<0.0001
	Hotellings						58.96	<0.0001
	Wilks						53.38	<0.0001
diagnosis level	Univariate: $df_h = 1$; $df_e = 14$	NA	21792.4	214488.7	21792.4	15320.6	1.42	0.25
		MOD	8.67	9.59	8.67	0.68	12.66	0.003
		SDOD	0.63	1.69	0.63	0.12	5.26	0.04
		SkewOD	130.17	145.75	130.17	10.41	12.50	0.003
		ExcOD	0.30	103.78	0.30	7.41	0.041	0.84
		MinOD	1.29	1.15	1.29	0.082	15.65	0.001
		MaxOD	20.86	51.15	20.86	3.65	5.71	0.03
	Multivariate:							
	Pillais						1.68	0.24
	Hotellings						1.68	0.24
	Wilks						1.68	0.24
Model 2 (diagnosis vs separate cells, Fig. 1B)	Univariate: $df_h = 1$; $df_e = 3198$	NA	21792.4	799415.0	21792.4	249.97	87.2	<0.0001
		MOD	8.67	26.57	8.67	0.0083	1043.0	<0.0001
		SDOD	0.63	3.88	0.63	0.0012	521.8	<0.0001
		SkewOD	130.17	699.80	130.17	0.22	594.8	<0.0001
		ExcOD	0.30	1199.04	0.30	0.37	0.81	0.368
		MinOD	1.29	6.66	1.29	0.0021	618.1	<0.0001
		MaxOD	20.86	137.02	20.86	0.043	486.8	<0.0001
	Multivariate:							
	Pillais						165.8	<0.0001
	Hotellings						165.8	<0.0001
	Wilks						165.8	<0.0001
Model 3 (diagnosis vs average values, Fig. 1C)	Univariate: $df_h = 1$; $df_e = 14$	NA-M	108.96	1072.46	108.96	76.60	1.42	0.25
		MOD-M	0.043	0.048	0.043	0.0034	12.67	0.003
		SDOD-M	0.0032	0.0084	0.0032	0.0006	5.27	0.04
		SkewOD-M	0.65	0.73	0.65	0.052	12.51	0.003
		ExcOD-M	0.0015	0.52	0.0015	0.037	0.041	0.84
		MinOD-M	0.0064	0.0057	0.0064	0.0004	15.56	0.001
		MaxOD-M	0.10	0.26	0.10	0.018	5.68	0.03
	Multivariate:							
	Pillais						1.63	0.25
	Hotellings						1.63	0.25
	Wilks						1.63	0.25

Table 1
(Continued)

Models and effects	Tests	Variables	Hypoth. SS	Error SS	Hypoth. mean SS	Error mean SS	F	p
Model 4	Univariate:	Fac-1.1	2.76	12.24	2.76	0.87	3.16	0.097
(diagnosis vs	$df_h = 1$;	Fac-2.1	0.39	14.64	0.39	1.04	0.37	0.55
patients, after	$df_e = 14$	Fac-3.1	8.37	6.63	8.37	0.47	17.69	0.001
calculation of		Fac-1.2	2.40	12.60	2.40	0.90	2.66	0.13
additional		Fac-2.2	0.047	14.95	0.047	1.07	0.044	0.84
indices and		Fac-3.2	3.65	11.35	3.65	0.81	4.50	0.052
dimensionality		Fac-4.2	0.021	14.98	0.021	1.07	0.02	0.89
reduction)		Fac-1.3	1.79	13.21	1.79	0.94	1.90	0.19
		Fac-2.3	6.83	8.17	6.83	0.58	11.71	0.004
		Fac-3.3	1.31	13.69	1.31	0.98	1.34	0.27
	Multivariate:							
	Pillais						15.0	0.004
	Hotellings						15.0	0.004
	Wilks						15.0	0.004

Notes: SS, sum of squares; df_h, hypothesized degrees of freedom; df_e, degrees of freedom for error terms.

Table 2
Results of stepwise variable selection and discriminant function properties in approaches II and III

Approach	Stepwise selection		Discriminant function		
	Selected variables	F to remove	Wilks' lambda	χ^2	Significance
Approach II	MOD	531.0	0.754		
	MaxOD	45.0	0.747		
	NA	25.9	0.739		
	ExcOD	13.3	0.736	978.9	<0.00001
Approach III	Var-3.1	10.2	0.442		
	Var-2.3	5.9	0.305	15.5	0.0004

Table 3
Performance of classification models in approaches II and III after the "jack-knife" validation

Approach	Predicted diagnosis	True diagnosis		Detection of malignancy	
		Adenoma	Carcinoma	Sensitivity	Specificity
Approach II	Adenoma	6	0	100%	75%
	Carcinoma	2	8		
Approach III	Adenoma	7	1	87.5%	100%
	Carcinoma	0	8		

4.1. Approach I

Since this approach accounts exactly for the innately nested structure of karyometric data, it would by far be the best way of statistical analysis in morphometry. Unfortunately, the range of statistical methods, which have been adopted to the nested design, is to date extremely narrow. In fact, the only well-established procedure, which completely fits this purpose is an ANOVA with mixed effects, or the variance component model [6,13,19], but even this is often absent in popular statistical packages [19]. As for other procedures, including DA and similar techniques, they were extended to the nested design only recently. In periodontal research, for example, methods such as generalized estimating equations, weighted least-squares analysis or regression models for correlated observations have been suggested [16,33]. There is also an entire class of multilevel statistical models, developed precisely for analyzing hierarchical data [26]. Theoretically, it is possible to fit discriminatory multilevel models which allow classification of both tumours and

cells [26]. In practice, however, none of the methods mentioned can be utilized by non-professionals, especially when classification rules are concerned. A further adaptation of these techniques for morphometric research is necessary.

It is probably this lack of easy-to-use methods that led to a wide implementation of two other approaches (approach II and approach III) in karyometry. In essence, both of them represent an attempt to simplify the nested design, i.e., to convert it into a factorial one, by removing either patients (approach II) or cells (approach III) from the analysis. It is, however, very important to know prior to such a conversion, how the precision and correctness of statistical analysis will be influenced.

4.2. Approach II

Table 1 clearly shows that model 2, in comparison with model 1, yielded strikingly lower p-values. It can be stated that this increase in significance is artificial and occurs due to a violation of one of the basic assumptions required by usual statistical methods, namely the assumption of independence of analytical units. An explanation for this statement, borrowed mainly from similar studies in periodontology, follows below.

As mentioned earlier, a nested design in morphometry occurs as a natural consequence of the corresponding sampling procedure. This procedure can be defined as a two-stage cluster sampling [42], where tumours represent naturally occurring clusters of cells. The natural origin of clusters here implies that their members tend to be rather alike. This similarity can be expressed quantitatively by an intraclass correlation coefficient (ICC) as follows [26,32,33]:

$$\mathrm{ICC} = \frac{\sigma_{\mathrm{p}}^2}{\sigma_{\mathrm{p}}^2 + \sigma_{\mathrm{c(p)}}^2}, \tag{1}$$

where σ_{p}^2 denotes the between-patient variance, and $\sigma_{\mathrm{c(p)}}^2$ the within-patient variance of cells; both these values are obtainable from the standard variance components table (see, for example, model 1 in Table 1, where hypothesis SS corresponds to σ_{p}^2 and error SS to $\sigma_{\mathrm{c(p)}}^2$). If ICC is greater than zero, the objects (cells) represent *dependent* observations [10,16,17,20,26,32, 33]. Under this condition, the traditional statistical formulas used in approach II underestimate the standard error of the mean [10,26,32]. The size of the underestimation is often referred to as a deflation factor (F_{d}) and can be calculated as follows:

Table 4
Intraclass correlation coefficients (ICC) and deflation factors (F_{d}) calculated for every karyometric feature

Parameter	ICC	True F_{d} (ratio F_1/F_2)
NA	0.27	61.3
MOD	0.36	82.4
SDOD	0.43	99.3
SkewOD	0.21	47.6
ExcOD	0.087	19.8
MinOD	0.17	39.5
MaxOD	0.37	85.3

$$F_{\mathrm{d}} = 1 + \mathrm{ICC}(k-1), \tag{2}$$

where k is the number of repeated measurements taken on each patient [32]. As a consequence, F-values in ANOVA, for instance, become F_{d} times as high as the true ones, which greatly reduces the corresponding p-levels [10,32]. We calculated ICC and F_{d} for each karyometric feature in our study[3] (see Table 4). The obtained ICC values were comparable with those in periodontology [20], but the F_{d} was always much higher due to the greater number of repeated measurements per patient [32].

The bias described has its consequences not only in ANOVA. In DA, for example, it leads to erroneous results in stepwise variable selection. As can be seen from Table 2, inclusion of each additional variable into analysis after MOD caused only a very slight decrease of the Wilk's lambda. In other words, the contribution of these variables to the discrimination between the tumour groups was extremely small, and yet they all entered the analysis due to the "highly significant", but, in fact, just inflated F-values. By analogy, the significance level of the discriminant function itself is also strongly biased.

A further question arising in this situation is to which extent our conclusions can be generalized, i.e., whether the error will always be so large on different tissue/tumour types. As can be seen from Eq. (2), F_{d} approximates 1 if either ICC or k is close to zero. However, in a morphometric study k should be at least 100,

[3]The F_{d} values given in Table 4 are true and correspond exactly to the ratio between F values in ANOVA-models 1 and 2. This is somewhat different from the results obtainable by using the Eq. (2), because the latter had been pruned to some extent [32]. For a precise quantification of F_{d}, not k but the corresponding degrees of freedom for cell and patient levels should be involved.

otherwise cell samples are not representative of their tumours [23]. As for ICC, its magnitude presumably may vary across different tissue types. It can be suggested from the analysis of Eq. (1) that ICC will be relatively large, for example, in highly differentiated neoplasms (such as those in our study), due to a low intra-tumour variability of cells ($\sigma^2_{\mathrm{c(p)}}$). On the contrary, in poorly differentiated tumours with marked nuclear atypia, where σ^2_{p} is much larger than $\sigma^2_{\mathrm{c(p)}}$, ICC will probably decrease. However, ICC can never be equal to zero, because some patient-to-patient variability exists in any case [42], thus the denominator in Eq. (1) is always more than 0. Moreover, it is well known in statistics that this type of variability (between primary sampling units) is usually the largest [30]. Indeed, with only a few exceptions [7], the contribution of the patient level was highly significant not only in our study, but also in many other karyometric investigations, regardless of the type of material used [8,9,24,31]. But even if the contribution of the patient level has been found to be non significant and the ICC value is very low (say, 0.05), the resulting error will still be considerable (F_{d} of app. 6 at $k = 100$) due to the large number of cells required for measurement in each patient.

Thus, it can be concluded that hypothesis tests in approach II unavoidably produce falsely lowered p-values, and although the size of the error may, probably, vary from tissue to tissue, it can hardly be neglected. The problem in approach II is, however, much worse than simply inaccurate hypothesis testing. *The described bias will remain virtually the same in all common statistical methods even if they do not involve hypothesis testing at all (e.g., factor or principal component analysis), because they all require the analytical units to be statistically independent* and, in addition, imply very similar computational routines [4,32, 33,39]. We would like to emphasize that the primary cause of this problem is the disregard of the patient as an analytical unit, which "...creates a target population statistically, that literally does not exist in nature" [33]. Because of this, an accurate statistical inference becomes *a priori* impossible, and it does not matter whether hypothesis testing was used or not [26]. Furthermore, the situation becomes especially piquant when, together with karyometric features, some other variables such as gender, age or tumour size must be considered. Indeed, all these parameters refer not to cells but to patients, i.e., to the higher level units [26], and it is unclear how they should be treated then.

4.3. Approach III

In this approach, the nested design is converted into a simple factorial one by averaging karyometric features across each tumour. In other words, nuclear parameters derived from separate cells are regarded as repeated measurements taken from corresponding objects (tumours). Such a definition appears to be quite sensible and, as the results of our study show, this approach produces rather accurate results: both models 1 and 3 yielded almost identical F- and p-values[4] (see Table 1). Evidently, this method of analysis leads to a correct statistical inference and, furthermore, allows the easy incorporation of any additional tumour and/or patient characteristics. Although, in comparison with approach I, it is not possible here to test the significance of *inter*-tumour variation, an *intra*-tumour variability of karyometric features (i.e., nuclear pleomorphism) can be evaluated instead. This is done by computing, in addition to the mean, several auxiliary distributional estimates [5]. In our study, these newly calculated variables turned out to be important for distinguishing between the tumour groups (see Table 1, model 4 vs. model 3). Thus, approach III also has an advantage of gaining some additional information, which can be of biological importance.

There are, however, two substantial disadvantages of this approach, which limit its use to some extent. The first problem is the usually small number of patients in each diagnostic group. The matter in question is not even the patient number itself, but rather the ratio of observations vs. variables (ROV, see above), which becomes particularly small when all the aforementioned distributional indices have been calculated. As a result, the covariance matrices in MANOVA become redundant, and discriminant function(s) computed in DA extremely unreliable [4,13,18]. Of course, the best solution would be to encompass more patients in the study, but in practice this often cannot be done, for example, due to the rarity of a disease [2]. Consequently, one usually has to lower the ROV from the opposite side, i.e., to abridge the number of dimensions (variables) in the analysis. It should be stressed at this point that no methods of variable screening based on statistical significance (e.g., forward or backward variable selection) are suitable for this purpose, because they involve a multiple comparison problem and, moreover, may lead to a substantial loss of practically impor-

[4] A slight difference in the multivariate tests occurred due to an approximation used when computing the mixed-effects ANOVA.

tant information [18,28]. Other procedures, which do not utilize the outcome variable, must be used [28,36]. Marshal et al. [36] have found that the best method is a nearest-neighbor clustering of variables followed by principal component analysis with orthoblique rotation. We also used this technique (slightly modified) in our study and, indeed, achieved a three-times dimensionality reduction without any substantial loss of information. Only after that, a limited stepwise variable selection may be carried out [28]. For further discussion of methods of data reduction we refer the reader to a number of excellent articles [18,28,36].

The second problem is that with taking average values in every tumour, a unique information characterizing each cell as a separate vector of all karyometric features gets irretrievably lost. In this respect, a clear distinction should be made between two types of pathological material to be measured. The first type comprises virtually all histologic and some kinds of cytologic specimens (for example, fine needle thyroid aspirates), where all cells selected for measurement belong to one and the same class (e.g., either benign or malignant). In this case, approach III is appropriate, as also shown by the results of our study. The second type of material comprises a major part of cytological preparations (exfoliative cytology, pleural and abdominal effusions), where varying proportions of cells of different types (from normal through dysplastic to malignant) may be present. Typically, the cells of interest (usually malignant) are rather scant here, but, ideally, even a single such cell should be reliably detected [7,9, 31,43]. Under this condition, taking average values can obscure biologically important events, so approach III is not encouraged in this situation.

4.4. Approach II vs. approach III

As discussed above, the ANOVA-model in approach II produced wrong significance levels; on the contrary, approach-III supplied accurate results. Apart from hypothesis testing, however, it is also interesting to compare these approaches with respect to the classification models developed. As our results show (see Fig. 4 and Table 3), approach III tends to produce more confident and reliable grouping of the tumours than approach II. It should be noted, however, that an exact evaluation of the model performance requires a larger sample size and stricter evaluation techniques, such as "hold-out" method or bootstrapping [18,28]. Furthermore, creation of a classification model is, in contrast to ANOVA, a complicated interactive process depending both on statistical settings and biological properties of the material. It is therefore difficult to predict whether changing tissue type will always influence both approaches to the same degree. Thus, our finding may not necessarily hold in all instances and requires further study.

Table 5
Summary of the features of investigated approaches and suggestions for use

Approach	Advantages	Disadvantages	Suggestions for use
Approach I	• Treats morphometric data as they are, i.e., hierarchically organized • Allows to evaluation of inter-patient variability of karyometric features • Potentially allows classification of both tumours and cells	• To date, classification models and many other statistical methods are not well established and difficult to apply	Can be recommended for universal usage in morphometry; further developments in the direction of classification modeling are necessary
Approach II	• Permits the classification of each separate cell • Is applicable to any type of pathological material	• Leads to a highly biased output in all statistical procedures	Should be avoided. If used, the imprecision of results must be clearly addressed
Approach III	• Permits the evaluation of intra-tumour variability of karyometric features (nuclear pleomorphism) • Permits the use of all traditional statistical methods, including classification modeling	• Is not applicable to many types of cytologic material • Can result in an excessive dimensionality and critical ROV-values	Recommended for use, if applicable. After computation of additional variables, a dimensionality reduction not involving the outcome variable is highly advisable

4.5. Conclusions and final remarks

The advantages and disadvantages of the approaches discussed above are summarized in Table 5. We also included in this table our suggestions concerning the use of these approaches with the hope that they could be useful for future morphometric investigations. On the whole, approach I seems to be the most appropriate for morphometry, but the corresponding statistical techniques are at present not well developed. If choosing between approaches II and III, the latter should be preferred wherever possible. The use of approach II must be restricted only to those situations, where approach III is inappropriate and there is no other way except to violate the assumption of independence. However, the researcher should be aware of the drawbacks of such analysis and "... consider their possible effects on the results and their interpretation (rather) than... ignore them in the hope that they will not be noticed" [1].

Finally, the practical meaning of the revealed differences between oxyphilic thyroid adenomas and carcinomas shall briefly be discussed. Certainly, the reclassification results obtained in the approach III (see Table 3) are very encouraging and, moreover, they are consistent with the results obtained in our previous series [44]. However, we are not inclined to make any definitive conclusions on this basis, mainly because the patient number was far too small in the present study, even with regard to the dimensionality reduction adopted. Besides, it is highly advisable to use more objective procedures for validation of multivariate models, as mentioned above. Thus our investigation should be considered in this respect as a feasibility study only. Further work is necessary to verify the efficacy and stability of the model created.

Acknowledgements

This work was supported by a grant from the Austrian Division of the International Academy of Pathology to O.T.

References

[1] D.G. Altman, S.M. Gore, M.J. Gardner and S.J. Pocock, Statistical guidelines for contributors to medical journals, *BMJ* **286** (1983), 1489–1493.

[2] D.G. Altman, *Practical Statistics for Medical Research*, Chapman & Hall, London, Glasgow, Weinheim, 1995.

[3] J.P.A. Baak, P.H.J. Kurver and M.E. Boon, Computer-aided application of quantitative microscopy in diagnostic pathology, *Path. Annu.* **17** (1982), 287–306.

[4] K. Backhaus, B. Erichson, W. Plinke and R. Weiber, *Multivariate Analysemethoden*, Springer, Berlin, Heidelberg, New York, 1996.

[5] P.H. Bartels, Numerical evaluation of cytologic data I, Description of profiles, *Anal. Quant. Cytol.* **1** (1979), 20–28.

[6] P.H. Bartels, Numerical evaluation of cytologic data XI, Nested designs in multivariate analysis of variance, *Anal. Quant. Cytol.* **4** (1982), 81–89.

[7] P.H. Bartels, Y.P. Chen, B.G. Durie, G.B. Olson, L. Vaught and S.E. Salmon, Discrimination between human T and B lymphocytes by computer analysis of digitized data from scanning microphotometry II, Discrimination and automated classification, *Acta Cytol.* **22** (1978), 530–537.

[8] M. Bibbo, P.H. Bartels, M. Salguero, H.E. Dytch, E. Lerma-Puertas and H. Galera-Davidson, Karyometric marker features in fine needle aspirates of microinvasive follicular carcinoma of the thyroid, *Anal. Quant. Cytol. Histol.* **12** (1990), 42–47.

[9] M. Bibbo, P.H. Bartels, M. Chen, M.J. Harris, B. Truttman and G.L. Wied, The numerical composition of cellular samples from the female reproductive tract I, Carcinoma in situ, *Acta Cytol.* **19** (1975), 438–447.

[10] N. Blomqvist, On the choice of computational unit in statistical analysis, *J. Clin. Periodontol.* **12** (1985), 873–876.

[11] A. Böcking, F. Giroud and A. Reith, Consensus report of the European Society for Analytical Cellular Pathology task force on standardization of diagnostic DNA image cytometry, *Anal. Quant. Cytol. Histol.* **17** (1995), 1–6.

[12] K.R. Castleman and B.S. White, The tradeoff of cell classifier error rates, *Cytometry* **1** (1980), 156–160.

[13] R. Christensen, *Analysis of Variance, Design and Regression*, Chapman & Hall, London, Weinheim, New York, 1996.

[14] CIRES, *Cell Image Retrieval and Evaluation System*, Kontron Elektronik GmbH, München, 1994.

[15] F. Collin, I. Salmon, I. Rahier, J.L. Pasteels, R. Heimann and R. Kiss, Quantitative nuclear cell image analyses of thyroid tumours from archival material, *Hum. Pathol.* **22** (1991), 191–196.

[16] T.A. DeRouen, Biostatistical and methodological issues in demonstrating efficacy of therapeutic agents for periodontal disease, *J. Dent. Res.* **68** (1989), 1661–1666.

[17] T.A. DeRouen, P.P. Hujoel and L.A. Mancl, Statistical issues in periodontal research, *J. Dent. Res.* **74** (1995), 1731–1737.

[18] *Panel on Discriminant Analysis, Classification and Clustering, Discriminant Analysis and Clustering*, National Academy Press, Washington, 1988.

[19] R.C. Elston and W.D. Johnson, *Essentials of Biostatistics*, F.A.Davis, Philadelphia, 1994.

[20] L.J. Emrich, Common problems with statistical aspects of periodontal research papers, *J. Periodontol.* **61** (1990), 206–208.

[21] O. Ferrer-Roca, E. Ballester-Guardia and J.A. Martin-Rodriguez, Morphometric, densitometric and flow cytometric criteria for the automated classification of thyroid lesions, *Anal. Quant. Cytol. Histol.* **12** (1990), 48–55.

[22] O. Ferrer-Rocca, E. Ballester-Guardia and J. Martin, Nuclear chromatin texture to differentiate follicular and papillary carcinoma of the thyroid, *Path. Res. Pract.* **185** (1989), 561–566.

[23] J.C. Fleege, P.J. van Diest and J.P.A. Baak, Computer assisted efficiency testing of different sampling methods for selective nuclear graphic tablet morphometry, *Lab. Invest.* **63** (1990), 270–275.

[24] H. Galera-Davidson, P.H. Bartels, A. Fernandez-Rodriguez, H.E. Dytch, E. Lerma-Puertas and M. Bibbo, Karyometric marker features in fine needle aspirates of invasive follicular carcinoma of the thyroid, *Anal. Quant. Cytol. Histol.* **12** (1990), 35–41.

[25] C. Giardina, L. Pollice, R. Ricco, E. Vacca, A. Penella, G. Serio, R. Mastroguilio, F. Potente and V.P. Delfino, Differential diagnosis between thyroid follicular adenoma and carcinoma: Analytic morphometric approach, *Path. Res. Pract.* **185** (1989), 726–728.

[26] H. Goldstein, *Multilevel Statistical Models*, Edward Arnold, London, Sydney, Auckland, 1995.

[27] P.W. Hamilton and D.C. Allen, Morphometry in histopathology, *J. Pathol.* **175** (1995), 369–379.

[28] F.E. Harrell, K.L. Lee and D.B. Mark, Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Stat. Med.* **15** (1996), 361–387.

[29] C. Hedinger, *Histological Typing of Thyroid Tumours*, Springer, Berlin, Heidelberg, New York, 1988.

[30] N.T. James, Common statistical errors in morphometry, *Path. Res. Pract.* **185** (1989), 764–768.

[31] L.G. Koss, P.H. Bartels, J.J. Sychra and G.L. Wied, Diagnostic cytologic sample profile in patients with bladder cancer using TICAS system, *Acta Cytol.* **22** (1978), 392–397.

[32] L.L. Laster, The effect of subsampling sites within patients, *J. Periodontal. Res.* **20** (1985), 91–96.

[33] L.L. Laster, Analysis of data from clinical studies of localized juvenile periodontitis, Discussion, *J. Clin. Periodontol.* **13** (1986), 476–480.

[34] F. Liautaud-Roger, J. Dufer, M. Pluot, M.J. Delisle and P. Coninx, Contribution of quantitative cytology to the cytological diagnosis of thyroid neoplasms, *Anticancer Research* **9** (1989), 231–234.

[35] F. Liautaud-Roger, J. Dufer, M.J. Delisle and P. Coninx, Thyroid neoplasms: Can we do any better with quantitative cytology?, *Anal. Quant. Cytol. Histol.* **14** (1992), 373–378.

[36] G. Marshall, F.L. Grover, W.G. Henderson and K.E. Hammermeister, Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery, *Stat. Med.* **13** (1994), 1501–1511.

[37] R. Montironi, R. Alberti, S. Sisti, A. Braccischi, M. Scarpelli, and G.M. Mariuzzi, Discrimination between follicular adenoma and follicular carcinoma of the thyroid: preoperative validity of cytometry on aspiration smears, *Appl. Pathol.* **7** (1989), 367–374.

[38] R. Nafe, R.S. Fritsch, B. Soudah, A. Hamann and H. Choritz, Histomorphometry in paraffin sections of thyroid tumours, *Path. Res. Pract.* **188** (1992), 1042–1048.

[39] M.J. Norusis, *SPSS Professional Statistics 6.1*, SPSS Inc., Chicago, 1994.

[40] I. Salmon, P. Gasperin, J.L. Pasteels, R. Heimann and R. Kiss, Relationship between histopathologic typing and morphonuclear assessments of 238 thyroid lesions, *Am. J. Clin. Pathol.* **97** (1992), 776–786.

[41] I. Sassi, F. Mangili, M. Sironi, M. Freschi and A. Cantaboni, Morphometric evaluation of fine needle biopsy of single thyroid nodules, *Path. Res. Pract.* **185** (1989), 722–725.

[42] A. Stuart, *The Ideas of Sampling*, MacMillan, New York, 1984.

[43] L.D. True, Morphometric applications in anatomic pathology, *Hum. Pathol.* **27** (1996), 450–467.

[44] O. Tsybrovskyy, I. Vassilenko, S. Mannweiler and M. Klimpfinger, Multivariate karyometric approach in differential diagnosis of follicular thyroid neoplasms: A study of 31 cases, *Virchows Arch.* **433** (1998), 135–143.

[45] R.G. Wright, H. Castles and R.H. Mortimer, Morphometric analysis of thyroid cell aspirates, *J. Clin. Pathol.* **40** (1987), 433–445.