Application of multilevel models to morphometric data. Part 1. Linear models and hypothesis testing

O. Tsybrovskyy [a,*] and A. Berghold [b]

[a] *Department of Pathology, School of Medicine, University of Graz, Austria*

[b] *Institute for Medical Informatics, Statistics and Documentation, University of Graz, Austria*

Received 17 October 2002
Accepted 28 April 2003

Abstract. Morphometric data usually have a hierarchical structure (i.e., cells are nested within patients), which should be taken into consideration in the analysis. In the recent years, special methods of handling hierarchical data, called multilevel models (MM), as well as corresponding software have received considerable development. However, there has been no application of these methods to morphometric data yet. In this paper we report our first experience of analyzing karyometric data by means of MLwiN – a dedicated program for multilevel modeling. Our data were obtained from 34 follicular adenomas and 44 follicular carcinomas of the thyroid. We show examples of fitting and interpreting MM of different complexity, and draw a number of interesting conclusions about the differences in nuclear morphology between follicular thyroid adenomas and carcinomas. We also demonstrate substantial advantages of multilevel models over conventional, single-level statistics, which have been adopted previously to analyze karyometric data. In addition, some theoretical issues related to MM as well as major statistical software for MM are briefly reviewed.

Keywords: Image analysis, computer assisted, morphometry, statistical analysis, multilevel models, thyroid, follicular neoplasms

1. Introduction

Morphometric data usually derive from a so-called multistage sampling, or sub-sampling [31]. In the simplest example, one takes a random sample of patients (tumors) and then randomly sub-samples cells (nuclei) within each patient.[1] A hierarchical, multilevel data structure ensues, with cells (level-1 units) being "nested" within patients (level-2 units). The importance of this feature for the morphometric research was recognized already in seventies [28] and further discussed in early eighties [13–15]. This discussion was however restricted to a simple decomposition of the total variance of one single morphometric feature into different components attributed to each level of hierarchy. A possibility was also shown to incorporate some factors into the analysis by using nested analysis of variance (ANOVA), for example, to explore differences between nuclear features in benign and malignant lesions [2]. However, due to many limitations of classical nested ANOVA and lack of other methods of analysis, two other approaches to data analysis have commonly been used in the morphometry, as described in our previous paper [33]. The first approach ("pooling" method) treats cells as independent units of analysis, that is to say cells from different patients are pooled, e.g., into benign and malignant "cell populations". In the second approach (summary statistics method), morphometric data are summarized within each patient by calculating mean values, standard deviations, etc., which the subsequent statistical analy-

[*] Corresponding author: Dr. Oleksiy Tsybrovskyy, Department of Pathology, University of Graz, Auenbruggerplatz 25, 8036 Graz, Austria. Tel.: +43 316 385 80 461, +43 316 380 44 11; Fax: +43 316 385 34 32, +43 316 38 43 29; E-mail: oleksiy.tsybrovskyy@kfunigraz.ac.at.

[1] In consistency with our previous article [33], we will continue to use the words "cell" and "nucleus" as well as "patient" and "tumor" interchangeably throughout this paper, too, since their meanings in the present context are identical.

sis is performed on. The aim of both approaches is to eliminate the hierarchy of the data and make use of traditional, "single level" statistical methods. However, both approaches, and especially the "pooling" method, have certain disadvantages and cannot be universally recommended [33].

In recent years, considerable progress in managing multilevel data has been achieved. In particular, an entire class of regression models, most widely known as multilevel models (MM), has been specially developed to handle hierarchical data structures [4,12,18,25,29, 32]. A number of user-friendly and well-documented statistical programs for MM are now also available, and overwhelmingly increasing number of examples of using MM can be found, mainly in epidemiological and socio-medical research [3,7–9,21]. However, MM have not yet been used in the morphometry.

In this paper we report our first experience with MM in the analysis of morphometric data. The present study is a logical continuation of our previous work [33]. In Part 1, we focus on linear models, i.e., models where morphometric nuclear features serve principally as dependent variables. A comparison with traditional, single-level statistics is made in order to demonstrate the key advantages of MM. In addition, some theoretical issues related to MM as well as major statistical software for MM are briefly reviewed.

2. Materials and methods

Seventy-eight follicular neoplasms of the thyroid (34 adenomas and 44 carcinomas, archival material) were analyzed. The diagnoses were based on WHO criteria [17]. All specimens were fixed in buffered 10% formalin and conventionally embedded in paraplast. For karyometry, monolayer preparations of cell nuclei were made according to the method of van Driel Kulker et al. [34]. Briefly, 50 μm thick sections were cut in each neoplasm from a block with representative tumor areas, deparaffinized in xylene, rehydrated and incubated in 0.05% pronase solution for 30 minutes at 37°C with intermittent vortex mixing. After filtration through 50-μm nylon gauge, centrifugation and re-suspension of the sediment in 2% carbowax solution (MW 1500, Merck), monolayer preparations were obtained using a Shandon cytocentrifuge at 2000 rpm for 15 min. The slides were air-dried and stored at 4°C. Nuclei were stained after Feulgen in modification thionine-SO_2, using CAS (Becton Dickinson) staining kits and protocols. Prior to staining, carbowax was dissolved in distilled water for 30 min. Measurements were performed by means of a semiautomatic system for image analysis, composed of Eclipse E600 microscope (Nikon, Japan) with 100/1.4 Plan Apochromat oil immersion objective used for measurements, 3CCD video camera DXC-930P (SONY, Japan), IntriguePro image frame grabber (Integral Tech., USA), and personal computer (PII, 266 MHz, 64 MB RAM). The system was controlled by Optimas 6.5 image analysis software, with customized macros written in ALI (Analytical Language of Images) to automate karyometric measurements. A total of 8 features (see Table 1) were identified for 147 to 258 (200 on average) nuclei in each slide. Nuclear contours were determined automatically using a customized combination of two autothresholding algorithms ("minimize variance" and "search for minimum") implemented in Optimas [1]. The system was calibrated with a micrometrical scale in order to obtain NA values in μm^2. Densitometric and texture features were measured in arbitrary units based on gray levels, and performed according to the recent consensus report [16]. Calibration of illumination as well as background and glare correction was used to avoid any artificial changes in object brightness. For the glare correction we used our own, improved algorithm of pixel-based *pre*-correction (not published).

Table 1

Nuclear features measured, with corresponding abbreviations

Class of features	Feature	Abbreviation
Geometric	Nucleus area	NA
	Nucleus circularity [1]	NC
Densitometric	Mean gray value within a nucleus*	MGV
	Standard deviation of gray values within a nucleus*	SDGV
	Skewness of gray values within a nucleus*	SkewGV
	Kurtosis of gray values within a nucleus*	KurtGV
	Integrated optical density within a nucleus	IOD
Textural	Surface area density (chromatin coarseness)	SAD

Features marked by asterisk* were computed based on the gray values, not on the optical density values, because they much better approximate to the normal distribution. SAD is equivalent to nuclear surface area density as defined in [20] divided by the number of pixels within the nucleus.

3. Statistical analysis and results

3.1. Some theoretical considerations and multilevel software

Discussion of the multilevel theory is beyond the scope of this paper. For comprehensive review of the topic we refer the reader to a number of books [12, 18,29] and journal articles [4,25,32]. Here, we just address some main issues. We will consider 2-level models only.

First of all, it has to be stressed that we are speaking about *modeling* morphometric data. It means that we want to explain the level and variation of a certain nuclear feature (outcome) in dependence of some factors (effects, predictors, covariates). The background behind such modeling is to get a biological understanding of certain lesions. No classificatory models, i.e. discriminating between benign and malignant conditions are considered here. This is currently the topic of our ongoing research, and the results will be published in the near future.

Two kinds of effects exist in statistical modeling: fixed and random. An effect is fixed, if it affects the population mean, and random, if it affects the variance of the dependent variable. Fixed effects are represented in the model by coefficients (intercept and slopes), and random effects by corresponding variance components attributed to these effects. Note that a variable can represent both fixed and random effects simultaneously. For example, a very common hypothesis is that malignant tumors have, on average, larger nuclei in comparison with benign. Tumor dignity is considered a fixed effect here. However, one can also hypothesize that malignant tumors demonstrate more pronounced anisokaryosis than do benign tumors. That is, tumor dignity might also affect the variance of the outcome variable, which is a random effect. Corresponding specification of the model is necessary to account for both effects.

There are also variables that represent purely random effects. In particular, subject effects associated with the sub-sampling in the morphometry contribute only to the variance of nuclear features and thus are random [30,37]. That means that prior to adding any covariates in the model, the hierarchy of the data must be accounted for by obligatory inclusion of a random effect representing level-2 (e.g., patient's ID). As a result, the between-patient variance is separated form the remaining, within-patient variance.

This is essentially the way that MM work: the variance of the dependent variable is split into components corresponding to the levels of hierarchy. Thus, even the "empty" two-level model, i.e. a model, with just a constant term and "no" additional covariates, actually does include a covariate identifying level-2 units. Having this basic structure, one can add other covariates to account for some effects that are either fixed or random at different levels of hierarchy. Note that assumptions, residuals, and interpretation of the results are all level-related in multilevel modeling, which may offer additional challenges. Some further issues (variable interactions, correlated errors from different levels) can cause MM to be extremely complex. While presenting our results, we will step through the major types of MM in the ascending complexity, in order to show which information can be gained from them and how the model parameters should be interpreted.

There are different methods of estimation model parameters, which can roughly be subdivided into simulation and non-simulation techniques.

Non-simulation techniques:

- IGLS (iterated generalized least squares),
- RIGLS (restricted iterative generalized least squares),
- ML (maximum likelihood),
- REML (restricted maximum likelihood),
- MINQUE (Minimum Norm Quadratic Unbiased Estimator).

Simulation techniques:

- Parametric bootstrap,
- Non-parametric bootstrap,
- Monte Carlo–Markov chain (MCMC).

For normal responses, IGLS is equivalent to ML and RIGLS is equivalent to REML. RIGLS and REML produce slightly better variance estimates than IGLS and ML [12,25,29]. Simulation methods are more precise, especially as far as random parameters are concerned [12,25], but they are also much slower and computationally intensive. In our study, we used RIGLS estimation.

Like all linear models, linear MM have assumptions of normality, linearity and homoscedasticity [4,12,18, 25,29,32]. All these assumptions apply to every level of hierarchy and consequently must be checked for each level separately. In addition, there are two assumptions specific to multilevel designs: independence of level-1 and level-2 residuals and no autocorrelation

Table 2
List of software for multilevel modeling or with some "multilevel" features

Software (latest version available)	Short description	Internet address
MLwiN, version 1.10.0007 (previously also MLn)	Dedicated statistical program for multilevel modeling	http://multilevel.ioe.ac.uk/
HLM, version 5.04	Dedicated statistical program for multilevel modeling	http://www.ssicentral.com/hlm/hlm.htm
MIXOR, version 2.0, MIXREG, version 1.2, MIXNO, version 1.0, MIXPREG, version 1.1	Suite of dedicated statistical programs for multilevel modeling	http://tigger.uic.edu/~hedeker/mix.html
WesVar, version 4; version 2.12 downloadable	Computes variance estimates in complex sampling designs, including multistage, stratified, and unequal probability samples	http://www.westat.com/wesvar/
VARCL	Variance component analysis for up to 9 levels of nesting	http://www.assess.com/Software/VARCL.htm
WinBUGS, version 1.3	Program for **B**ayesian inference **U**sing **G**ibbs **S**ampling, allowing some kinds of multilevel modeling	http://www.mrc-bsu.cam.ac.uk/bugs/
SAS/STAT (GENMOD, MIXED and NLMIXED procedures), version 8.2	General purpose statistical program with powerful multilevel features	http://www.sas.com/rnd/app/da/stat.html
S-Plus, version 6	General statistical program with multilevel features	http://www.insightful.com/products/product.aps?PID=3
Oswald, version 3.4	Suite of S-plus functions for analyzing longitudinal data	http://www.maths.lancs.ac.uk/Software/Oswald/
STATA, version 7.0	General statistical program with some multilevel features	http://www.stata.com/
SPSS, version 11.0 (Advanced Models)	General statistical program with some multilevel features	http://www.spssscience.com/spss11/
SYSTAT 10.2	General statistical program with some multilevel features	http://www.systat.com/

at level-1 [25,32]. Normality, homoscedasticity, independence and autocorrelation of residuals can be inspected after a model is fitted using residual plots. MM allow an easy estimation and plotting of residuals for each level. Linearity assumption can be verified by including quadratic, cubic, etc. terms of corresponding covariates in the model.

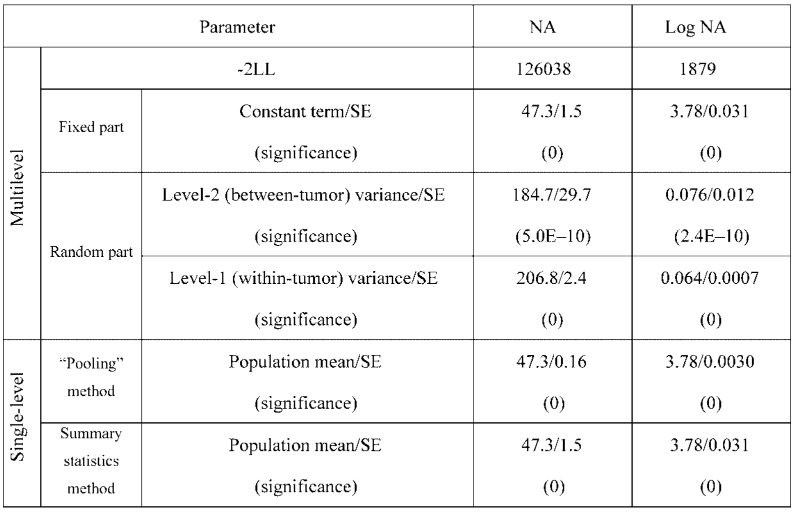
There are both general statistical packages with some multilevel procedures included and dedicated software for multilevel modeling. In Table 2, we listed all such programs that we are currently aware of. They differ considerably in the number of modeling possibilities, estimation methods and user-friendliness. Some of these packages have been reviewed in the literature in rather detail [6,30,36,37,39]. There is, however, no single program where all the estimation methods listed above would be implemented. To develop MM in our present work, we used MLwiN 1.10 [27], which presently seems to be the most advanced and, at the same time, very user-friendly program [6,36,39].

3.2. *Modeling hierarchy in "empty" model*

An "empty" two-level model includes a constant term (as a fixed effect) and two variance components corresponding to level-1 (within-tumor) and level-2

Table 3

"Empty" model for NA and log NA



Parameter			NA	Log NA
Multilevel	-2LL		126038	1879
	Fixed part	Constant term/SE (significance)	47.3/1.5 (0)	3.78/0.031 (0)
	Random part	Level-2 (between-tumor) variance/SE (significance)	184.7/29.7 (5.0E–10)	0.076/0.012 (2.4E–10)
		Level-1 (within-tumor) variance/SE (significance)	206.8/2.4 (0)	0.064/0.0007 (0)
Single-level	"Pooling" method	Population mean/SE (significance)	47.3/0.16 (0)	3.78/0.0030 (0)
	Summary statistics method	Population mean/SE (significance)	47.3/1.5 (0)	3.78/0.031 (0)

Notes: −2LL – negative doubled log likelihood. SE – standard error. Significance levels lower than 1.0E-99 are presented as 0.

Table 4

Most important summary from "empty" models and models with DIAGNOSIS as random coefficient covariate

Feature	"Empty" model		Model with DIAGNOSIS as a random coefficient covariate		
	Significance of hierarchy (level-2 variance)	ICC overall	DIAGNOSIS coefficient (significance)	Complex variance	
				Level-2: significance	Level-1: difference carcinomas – adenomas, % (significance)
Log NA	4.3E-10	0.54	0.22 (0.0001)	0.57	36% (1.9E-36)
NC	7.7E-10	0.26	0.84 (0.0078)	0.89	48% (1.4E-62)
MGV	4.8E-10	0.58	23.5 (1.4E-11)	1	−5% (0.018)
SDGV	6.0E-10	0.50	−0.89 (0.34)	0.73	−12% (8.8E-09)
SkewGV	4.4E-10	0.52	−0.44 (9.5E-11)	0.13	−31% (4.1E-57)
KurtGV	4.7E-10	0.52	−0.44 (0.00056)	5.9E-07	−37% (1.8E-87)
Log IOD	5.1E-10	0.51	−0.056 (0.19)	0.36	56% (9.7E-79)
SAD	5.2E-10	0.62	0.35 (4.5E-05)	0.89	53% (1.9E-74)

(between-tumor). In MLwiN, the model is specified by allowing the constant term to vary at both levels [27]. Note that level-2 variance represents the random effect associated with nesting (subject effect). The "empty" model is essentially equivalent to the variance component model available in all major statistical packages like SPSS or SAS [30,37]. It can be useful in three different ways (see Tables 3 and 4 for the results).

1. The most interesting feature for us is the possibility to explore the hierarchy itself – i.e., to judge how strong the clustering effects within the level-2 units are and whether we really do need to account for this. Table 3 shows a complete example of an "empty" model for the variable NA. As can be seen, the level-2 variance in this model is highly significant, which means the hierarchical structure of the data cannot be ignored. The between-patient variance was highly significant also for all other karyometric features measured (Table 4). This agrees with our results reported previously [33] and stresses once more the great importance of objects' hierarchy for analysis of morphometric data.

We can further get an idea about the magnitude of the clustering effects in our data – i.e., the degree of similarity of cells within each tumor. For this, we compute the intra-class correlation coefficient (ICC) using the known formula $\sigma^2_{\mathrm{p}}/(\sigma^2_{\mathrm{p}}+\sigma^2_{\mathrm{c(p)}})$ [10,12,29], where σ^2_{p} is the between-patient, and $\sigma^2_{\mathrm{c(p)}}$ the within-patient variance. As demonstrated in Table 4, the obtained

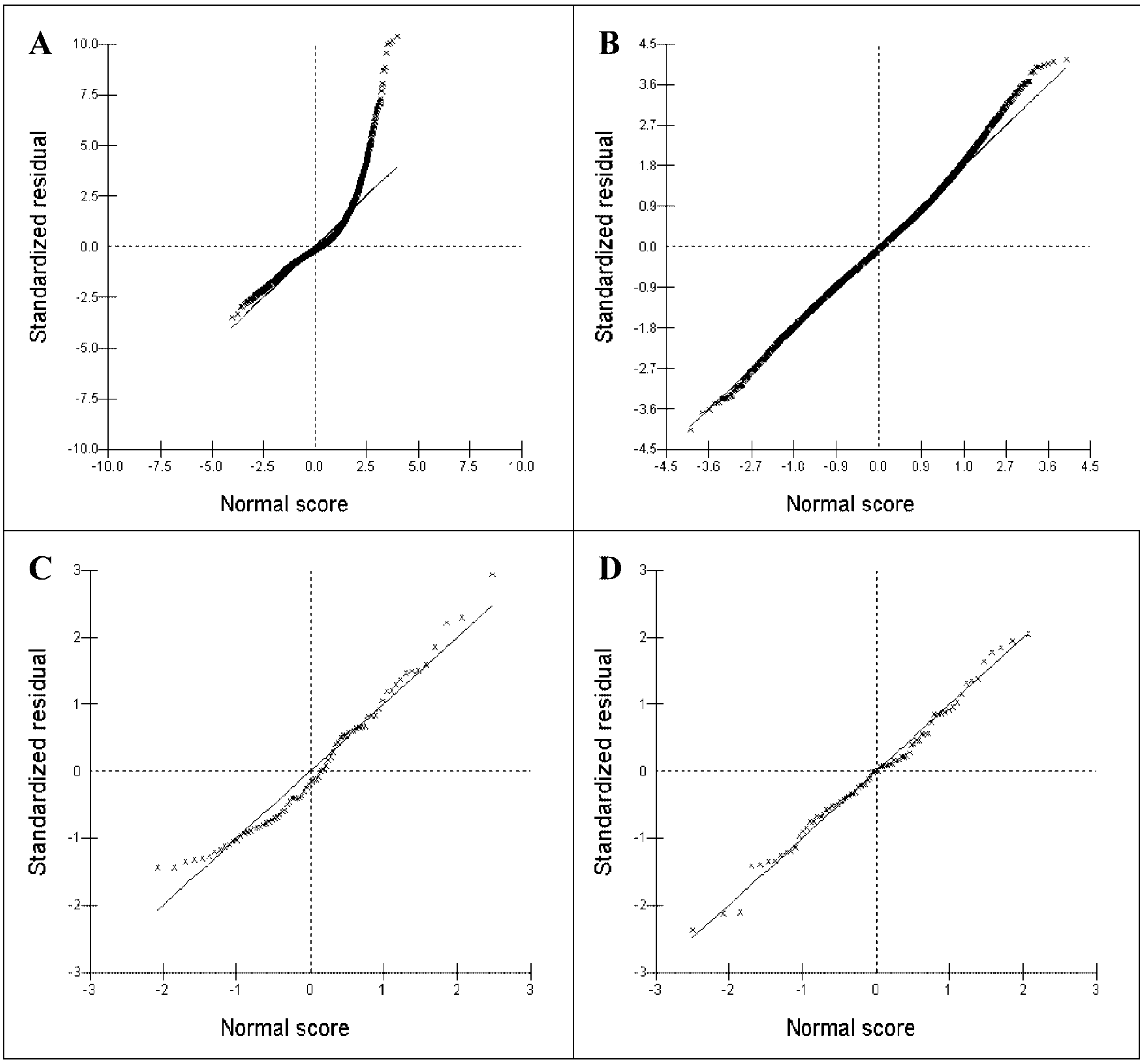


Fig. 1. Plots of residuals for the "empty" model of NA. A, B – level-1 residuals, before and after log transform, respectively. C, D – level-2 residuals, before and after log transform, respectively. Straight lines represent ideally normal distribution.

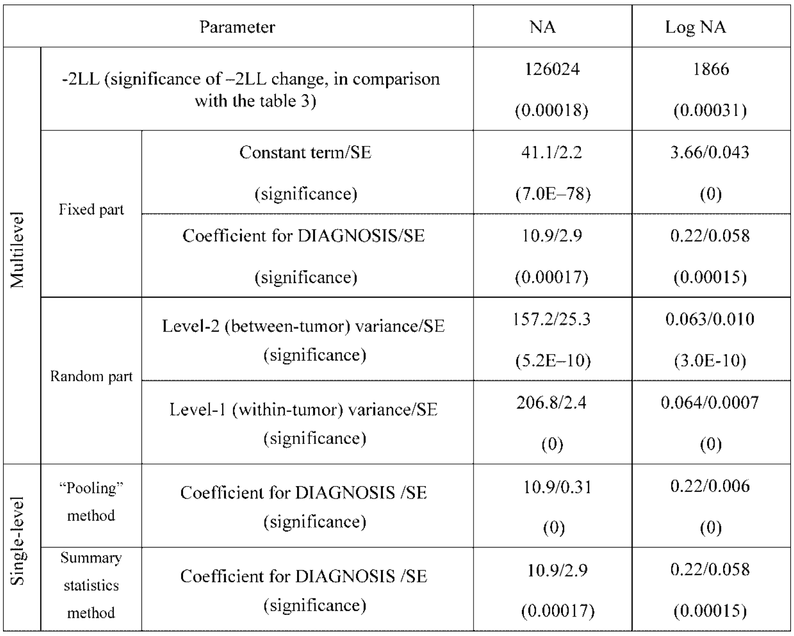
ICC values were mostly between 0.5 and 0.6, which is rather high – higher than our previous results for oxyphilic thyroid tumors [33].

2. Constant term in the "empty" model is interpreted as the mean NA value in the entire population of follicular thyroid tumors. Table 3 shows that mean NA computed in MM as well as in both single-level approaches was the same. However, the SE of mean in the "pooling" method was greatly (about 10-fold) underestimated, whereas in the summary statistics method it was correct. This fully agrees with our previous data [33].

3. The assumption of normality can be checked already in the "empty" model by examining plots of standardized residuals vs. normal scores. Figure 1A,C shows an example of NA, for which normality assumption was violated at both levels, especially at level-1. After log transform, only minor deviance from the normal distribution was seen (Fig. 1B,D). Note, however, that interpretation of the model coefficients after log transform is different. Constant term of 3.78 (see Table 3) means that the mean NA in the population is $e^{3.78} = 43.8\ \mu m^2$. This is lower than the mean estimated on the raw data, because the distribution of

Table 5

Model for NA and log NA with DIAGNOSIS as fixed coefficient covariate



Parameter			NA	Log NA
Multilevel	-2LL (significance of −2LL change, in comparison with the table 3)		126024 (0.00018)	1866 (0.00031)
	Fixed part	Constant term/SE (significance)	41.1/2.2 (7.0E−78)	3.66/0.043 (0)
		Coefficient for DIAGNOSIS/SE (significance)	10.9/2.9 (0.00017)	0.22/0.058 (0.00015)
	Random part	Level-2 (between-tumor) variance/SE (significance)	157.2/25.3 (5.2E−10)	0.063/0.010 (3.0E-10)
		Level-1 (within-tumor) variance/SE (significance)	206.8/2.4 (0)	0.064/0.0007 (0)
Single-level	"Pooling" method	Coefficient for DIAGNOSIS /SE (significance)	10.9/0.31 (0)	0.22/0.006 (0)
	Summary statistics method	Coefficient for DIAGNOSIS /SE (significance)	10.9/2.9 (0.00017)	0.22/0.058 (0.00015)

Notes: −2LL – negative doubled log likelihood. SE – standard error. Significance levels lower than 1.0E-99 are presented as 0.

NA is left-skewed. Note also that the relation between level-2 and level-1 variance has changed, and ICC increased from 0.47 on the raw data $(184.7/(184.7 + 206.8) = 0.47)$ to 0.54 on the log scale (see Table 4).

Violation of normality assumption was also detected for IOD, so the log transform was required here, too. All other nuclear features showed nearly normal distribution.

3.3. Exploring simple research hypothesis

Now we can extend the "empty model" to explore some research hypotheses. In our particular example, let us formulate the question of interest as that about magnitude and significance of the differences in nuclear features between follicular adenomas and carcinomas. Thus, we add a new term, DIAGNOSIS (coded as 0 for adenomas and 1 for carcinomas) as a fixed effect to the "empty" model created at the previous step. The full example of the model for NA is shown in Table 5, and the results are interpreted as follows.

1. −2 Log Likelihood (−2LL), in comparison with the empty model, significantly decreased (by 13) – i.e., the model fits the data better now.

2. Constant term shows the value of NA at 0 value of DIAGNOSIS. Since in our data set adenomas were coded as 0 and carcinomas as 1, the value of 41.1 corresponds now to the mean NA for *adenomas only.* For log transformed data, the estimated average NA is $e^{3.66} = 38.86\ \mu m^2$.

3. The coefficient for DIAGNOSIS indicates that nuclei in carcinomas are, on average, 10.9 μm^2 larger than in adenomas, and thus, the mean NA in carcinomas is $41.1\ \mu m^2 + 10.9\ \mu m^2 = 52.0\ \mu m^2$. As for log transformed data, the coefficient for DIAGNOSIS means that NA in carcinomas is, on average, $e^{0.22} = 1.246$ times as high as that in adenomas and is, thus, 48.42 μm^2. Note that, again, both single-level approaches yield correct mean values, but the standard error in the "pooling" method is greatly underestimated, leading to an extreme inflation of the significance value for DIAGNOSIS.

4. Level-2 variance remained highly significant, but, in comparison with the "empty" model, considerably decreased. By contrast, level-1 variance remained exactly the same. This is due to the fact that DIAGNOSIS is uniform within each level-2 unit and is, therefore, a level-2 covariate. In some other situations, e.g., while studying gynecologic smears with a mixture of benign and malignant cells on each slide, DIAGNOSIS can be level-1 covariate, so *both* level-1 and level-2 variances would change.

Table 6

Model for log NA with DIAGNOSIS as random coefficient covariate

Parameter			Constant term (intercept)	DIAGNOSIS (slope)
-2LL (Significance of -2LL change, in comparison with table 5)			1691 (1.0E-38)	
Fixed part		Coefficient/SE (significance)	3.66/0.041 (0)	0.22/0.057 (0.00011)
Random part	Level-2 (between-tumors)	Variance/SE (significance)	0.057/0.014 (4.7E–05)	0* (–)
		Covariance/SE (significance)	0.0056/0.010 (0.58)	
	Level-1 (within-tumors)	Variance/SE (significance)	0.053/0.0009 (0)	0* (–)
		Covariance/SE (significance)	0.0096/0.0007 (8.3E–43)	

Notes: −2LL – negative doubled log likelihood. SE – standard error. Parameters marked by asterisk * were pre-constrained to 0, because they were redundant [27]. Significance levels lower than 1.0E-99 are presented as 0.

3.4. Complex variance structure

Until now, the coefficient for DIAGNOSIS was modeled as fixed. This implies that both between- and within-tumor variance is the same for adenomas and carcinomas. It is, however, logical to hypothesize that due to usually higher cellular atypia in malignant tumors

(a) variation of nuclear size (and also other nuclear features) might be higher within carcinomas than within adenomas, and
(b) carcinomas might differ more from each other in their (average, predominant) nuclear size (and other nuclear characteristics) than do adenomas.

To check it, we allow the coefficient for DIAGNOSIS to vary at both levels. In other words, we consider now DIAGNOSIS to represent not only a fixed effect, but also two random effects: one corresponding to the question (a) above and affecting level-1 variance, and another one corresponding to the question (b) and thus affecting level-2 variance. Using terminology common to multilevel modeling, the coefficient for DIAGNOSIS is specified as *random*[2] at both levels [12, 29,27]. Due to additional random effects, a complex variance structure at each level results (see below). Table 6 shows model parameters for log NA. The interpretation is as follows:

1. −2LL decreased drastically (by 175, at 2 additional degrees of freedom), which means a great improvement of the model fit.

2. The fixed model coefficients (intercept and slope) remained practically unchanged, and their meaning is the same as in the previous step.

3. At each of the levels, we have 3 parameters now instead of 1: intercept variance, slope variance (variance due to the random effects of DIAGNOSIS) and covariance of intercept and slope. This allows us to compute the variance of NA separately for adenomas and carcinomas. Note that for this purpose, any two out of the three parameters are sufficient. This is why the variance of the slope (DIAGNOSIS) was pre-constrained to 0 [27]. The variance for adenomas has already been computed: it is the variance of the intercept. The variance for carcinomas is the variance for intercept plus twice the covariance between intercept and slope and is, thus, 0.068 for level-2 and 0.072 for level-1. Note, however, that the level-2 covariance is non-significant; the level-1 covariance is, by contrast, highly significant. We can thus state that the within-tumor variation of NA (i.e., nuclear polymorphism) is significantly higher in carcinomas than in adenomas.

[2] The term “random coefficient” is often used in text books on multilevel modeling [12,29] and designates, in fact, a variable representing an admixture of a fixed effect and at least one random effect at any of the levels.

For the between-tumor variance of NA, this tendency is also observed, but is not significant.

The summary of the models with complex variance for all nuclear features measured is presented in Table 4. Differences in mean values between adenomas and carcinomas were not significant for SDGV and IOD and highly significant for all other nuclear features. Complex variance at level-2 was not significant for all features, except KurtGV. At level-1, however, the situation was completely different. As can be seen, carcinomas showed much higher within-tumor variation of nuclear size and form factor, DNA amount and chromatin coarseness, and lower variation in nuclear staining intensity and its derivates.

Having separate variances for adenomas and carcinomas, we can further compute separate ICC for each tumor group. We can even construct confidence intervals for these ICC values, and thus compare them. In our study, however, ICC were not significantly different between carcinomas and adenomas, mostly due non-significant covariances at level-2, with an exception being, again, KurtGV (data not shown).

3.5. *Exploring complex research hypothesis*

In a similar way to the previous steps we can incorporate in the model some further covariates, which may be relevant to our research question. Again, their coefficients can be modeled either as fixed or random at each of the levels, depending on the nature of the features and the interpretability of the model parameters. As an example, we present in Table 7 a model for NA as dependent variable, and DIAGNOSIS and IOD as covariates with random coefficients at both levels. The reason for constructing such a model is the hypothesis that changes of NA might be secondary to the changes in nuclear DNA amount. Thus, we want to explore the difference of NA between carcinomas and adenomas while controlling for IOD. Both NA and IOD were log transformed to meet the assumption of normality. Furthermore, log IOD was then centered (by subtraction) around the value of 2.48, which corresponded to the nearly-diploid peaks on the DNA histograms. This greatly improved both convergence of the model and interpretability of the results.

1. The dramatic change of -2LL (by 8630 at 7 degrees of freedom) indicates a crucial improvement of model fit and confirms our surmise that NA is closely related to the nuclear DNA amount.

2. Constant term corresponds to the NA at 0 value of DIAGNOSIS (i.e., adenomas) and 0 value of log IOD (i.e., nearly-diploid DNA amount, due to the centering) and is equal to $e^{3.60} = 36.6\ \mu\text{m}^2$.

3. The coefficient for DIAGNOSIS shows that, *for the same DNA amount*, malignant nuclei are $e^{0.26} = 1.3$ times as large as benign nuclei. Thus, nearly-euploid nuclei in carcinomas are, on average, $36.6\ \mu\text{m}^2 \times 1.3 = 47.6\ \mu\text{m}^2$ in size. Note that, in

Table 7

Model for log NA with DIAGNOSIS and IOD as random coefficient covariates

Parameter			Constant term	DIAGNOSIS	Log IOD
-2LL (Significance of -2LL change, in comparison with table 6)			–6939 (0)		
Fixed part		Coefficient/SE (significance)	3.60/0.028 (0)	0.26/0.040 (8.0E–11)	0.95/0.030 (0)
Random part	Level-2 (between-tumors)	Variance/SE (significance)	0.026/0.0064 (4.9E–05)	0* (–)	0.059/0.011 (8.7E–08)
		Covariance/SE (significance)	0.0057/0.052 (0.91)	–0.0015/0.0073 (0.84)	0.0070/0.011 (0.52)
	Level-1 (within-tumors)	Variance/SE (significance)	0.033/0.00064 (0)	0* (–)	0.016/0.0043 (0.0002)
		Covariance/SE (significance)	0.0015/0.00043 (0.00047)	–0.0056/0.0015 (0.00019)	0.010/0.0018 (2.8E–08)

Notes: -2LL – negative doubled log likelihood. SE – standard error. Parameters marked by asterisk * were pre-constrained to 0, because they were redundant [29]. Significance levels lower than 1.0E10-99 are presented as 0.

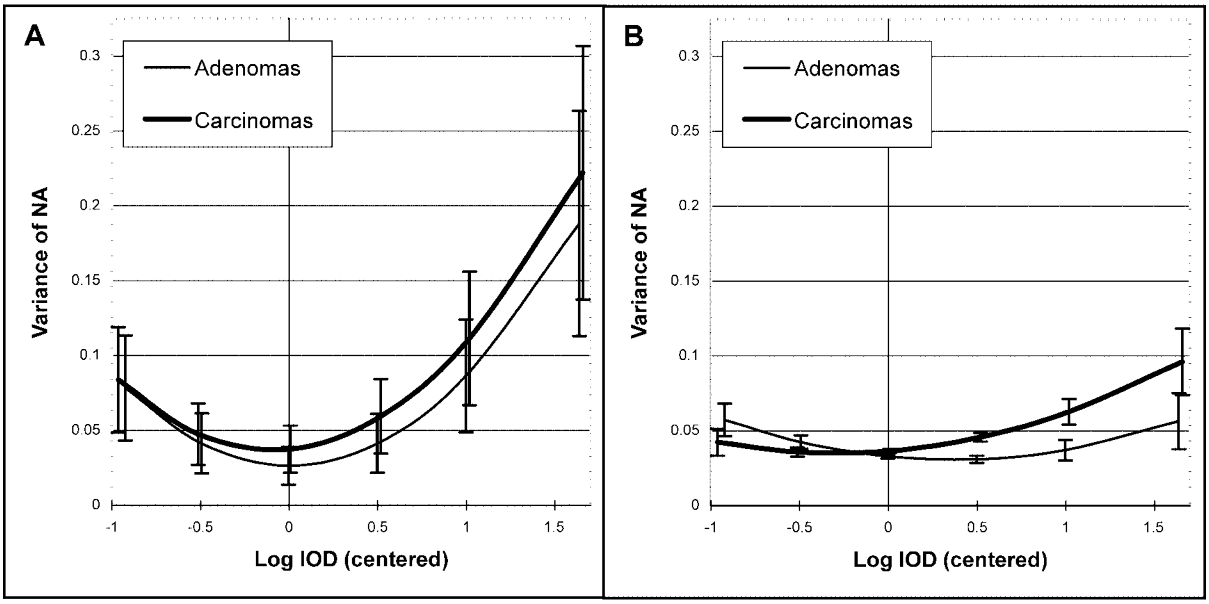


Fig. 2. Dependence of the variance of NA on IOD estimated separately for adenomas and carcinomas. A – between-tumor variance, B – within-tumor variance. Error bars represent 95% confidence intervals.

comparison with the previous model (Table 6), the coefficient for DIAGNOSIS increased, whereas its standard error remarkably decreased. This means that differences in nuclear size between adenomas and carcinomas became even more apparent if we control for nuclear DNA amount.

4. The coefficient for IOD indicates that NA is directly related to the nuclear DNA amount. The association function is of the power type; however, the coefficient is very close to 1, so the relationship is nearly linear, at least within the observed range of IOD values. For example, *within the same tumor type* (i.e., either adenomas or carcinomas), triploid nuclei are $e^{0.95\times\ln(1.5)} = 1.5^{0.95} = 1.47$, and tetraploid nuclei $e^{0.95\times\ln(2)} = 2^{0.95} = 1.93$ times as large as diploid nuclei.

5. The random parameters of the model allow us to evaluate the between- and within-tumor variance separately for adenomas and carcinomas in dependence on log IOD. This variance function is comfortably computed and plotted in MLwiN (see Fig. 2). As can be seen, the lowest NA variance corresponds to nearly diploid DNA amount, and the more is the deviance from the euploidy, the higher is NA variance. This is true for both levels, but is more pronounced at the level-2. Another point to note is that at high ploidy levels (beginning approximately from triploid), carcinomas show significantly higher nuclear pleomorphism (within-tumor variation of NA) than do adenomas (in Fig. 2B, the confidence limits do not overlap). Between-tumor variance seems to be equal for both tumor types, irrespective of nuclear DNA amount.

4. Discussion

In the present study, we could demonstrate some substantial advantages of MM over usual, single-level statistics in application to morphometric data. On the basis of the models fitted, we could draw a number of interesting conclusions:

(a) Data hierarchy in the morphometry is a very important factor, with ICC values ranging mainly between 0.5–0.6 (see Table 4), which is rather high – higher than in our previous study [33] and in other research fields with published ICC data [10].

(b) There are significant differences between follicular adenomas and carcinomas in nuclear size and form factor, staining intensity, and certain chromatin properties. In particular, malignant nuclei, in comparison with benign, are larger, more irregularly-shaped, stain paler, and have coarser chromatin. The magnitude of these differences is described by corresponding coefficients (see Table 4), and the example of their interpretation was given above.

(c) There seems to be no difference in nuclear DNA amount between adenomas and carcinomas. This finding is consistent with most other reports [5,23,26,38]. Furthermore, SDGV (which describes the variation of staining intensity across different points in a nucleus) appears also to be the same for both tumor types.

(d) Within-tumor variation of nuclear features is different for adenomas and carcinomas (see Table 4). Malignant tumors show higher nuclear pleomorphism, higher variation of nuclear DNA amount and higher variation of chromatin coarseness. By contrast, variation of nuclear staining intensity and its derivates is lower.[3] The latter finding is especially interesting with regard to the lower average staining intensity typical for malignant cells. It seems that malignant transformation affects *most* cells within the tumor in a uniform way in that they loose their tincture properties.

(e) Between-tumor variation of nuclear features, excepting KurtGV, appears to be the same for adenomas and carcinomas.

(f) NA is directly related to nuclear DNA content. When controlled for tumor dignity, this relationship is nearly linear – e.g., tetraploid tumors are almost 2 times as large as diploid. Note, however, that this holds for monolayer preparations used in our study, where nuclei are very much flattened. In paraffin sections, this relationship may be of other magnitude.

(g) Among cells with the same DNA content, malignant nuclei are about 1.3 times as large as benign nuclei.

(h) Diploid cells are the most uniform with regard to their nuclear size; at higher ploidy levels, nuclear pleomorphism (both between- and within-tumor) increases.

(i) Among cells with high DNA content (beginning approximately from triploid), malignant nuclei are more pleomorphic than benign.

Many of these conclusions would be impossible using single-level statistics, or we would end up with wrong conclusions at all. The "pooling" method leads to a large bias due to reasons considered in our previous work [33] and is, therefore, unacceptable. Summary statistics method is precise and effective enough for simple hypotheses, as can be seen on many examples in the literature [22,24,35]. However, it is ineffective for complex research questions [4]. In particular, estimation and comparison of within- and between-tumor variances is still possible (although very uncomfortable), if there is, at most, one *categorical* covariate (e.g., DIAGNOSIS). If we want to control for other, and especially continuous, covariates and explore the complex variance-covariance structure, like in our last model, MM are necessary.

The advantage of MM becomes even more apparent if there are more than 2 levels of nesting. For example, in a multicenter study, the centers (institutions) would represent the third level of the hierarchy. Or one can account for measurement error by taking repeated measurements from the same nuclei and treating them as the lowest-level units. It is clear that the "pooling" approach becomes even more imprecise, and summary statistics approach even more ineffective in this setting.

There are some extensions of traditional statistical methods, e.g., nested ANOVA and mixed-effect general linear models (GLM) capable of handling random effects. It is thus possible to manage certain kinds of multilevel data using GLM procedure available, for instance, in SPSS or SAS [30,37]. However, GLM produce estimates based on sums of squares and therefore achieve their optimal performance only on completely balanced designs. On unbalanced data, a bias occurs, requiring a careful choice of weighting [25,30]. GLM also treat all (even random!) effects as fixed and construct F statistics based on the ratio of the appropriate sums of squares. On the contrary, variance components, especially those attributed to random effects, are not estimated directly, and no standard errors are produced for them [30,37]. Consequently, it is in fact impossible to correctly test the significance of random effects using GLM. In contrast, MM use maximum likelihood estimation, which is asymptotically efficient even on unbalanced data [30]. Variance parameters are estimated directly in MM, together with corresponding standard errors [25,30]. Variance can also be modeled as a function of other covariates, like in our last model. This is impossible in GLM. In addition, MM generate -2 Log Likelihood and other indices like AIC (Akaike's Information Criterion) and BIC (Schwarz's Bayesian Information Criterion), which describe the model fit and can be used to search for "the best" model [12,29,30,37]. Finally, MM provide a natural way to handle multivariate outcomes [25], and we are going to demonstrate one of possible uses of these models in Part 2.

[3] It is known that statistical moments about the mean usually correlate with the absolute magnitude of the mean. However, the relationship observed was completely preserved even after SDGV, SkewGV and KurtGV were normalized by MGV (data not shown).

On the other hand, MM are much more complicated than conventional, single-level models. MM, due to their nature, split all variance components into different levels. Correspondingly, one has to check assumptions and interpret variance parameters separately for each of the levels. The general interpretation rule (on example of karyometric data) is that level-1 variance/covariance refers to cells, or “within-tumor” variation, and level-2 variance/covariance to tumors, or “between-tumor” variation of the dependent variable. If two or more covariates are present in the model, the meaning of the parameters for each variable becomes conditional on the presence of other variables in the model, which increases even more interpretation challenges. In addition, all the problems common to biostatistical modeling (e.g., sufficient number of study objects and adequate ratio “observations-to-variables”, multiple comparisons problem, selection of the “best” model, checking assumptions, etc.) are no less and often even more acute in multilevel designs.

A particular issue to be addressed is the robustness of multilevel estimates against small samples and violation of assumptions, especially the one of normality. When using MM, one should consider sample size at both levels, i.e. number of level-2 units and number of level-1 units per level-2 unit. In the morphometry, a sufficient number of level-1 units per level-2 unit (e.g., several hundreds of cells per tumor) is usually easy to achieve. It is also desirable to have more than 100 level-2 units in a sample, especially accounting for the great magnitude and importance of the between-subject variation [13–15,28,33]. Yet, quite often the number of level-2 units is rather small, for example, due to the rarity of a particular tumor. As a result, it is difficult or even impossible to rule out non-normality using plots of level-2 residuals. It is, however, reasonable to assume a normal distribution of level-2 variables due to the central limit theorem [11]. In addition, the most recent study by Maas and Hox [19] based both on a comprehensive literature review and own simulations shows that multilevel estimates are generally robust to moderate non-normality and small sample sizes. Samples with about 50 level-2 units produce unbiased estimates and standard errors, except for the standard errors of variance components at level-2 [19]. To eliminate this bias, bootstrap or simulation techniques can be effectively applied [12,25,27]. MLwiN offers very comfortable facilities for parametric and non-parametric bootstrap, Gibbs sampling and MCMC.

There are also some purely technical, computational problems related to MM and/or to the corresponding computer program. Quite frequently, we experienced numerical errors and even program crashes while running MLwiN, even if the model was specified correctly. This was particularly typical for models with complex variance structure. Newer versions of the software might remedy the problem. According to our experience, if severe numerical errors occur or coefficients do not converge, switching between different estimation options (IGLS/RIGLS, negative variances at each level allowed/not allowed) may solve the problem. In all situations, centering the data around the grand mean as well as transforming them in order to improve normality are helpful. However, the interpretation of the models fitted to transformed and/or centered data is different and somewhat more complicated, as demonstrated in our study.

In conclusion, MM represent a very powerful and flexible way for modeling and analyzing morphometric data. They are superior over other statistical approaches that have been used in the morphometry until now. However, MM are more complex than common single-level statistics, and require very careful and thorough specification and interpretation. We hope that our experience described in this paper will be useful for those morphometrists who are going to apply MM in their research. For systematic reading on MM, a number of sources can be recommended [4,12,18,25,29,32].

References

[1] *Optimas 6.5: User Guide and Technical Reference,* Media Cybernetics, Silver Spring, 1999.

[2] P.H. Bartels, Numerical evaluation of cytologic data XI. Nested designs in multivariate analysis of variance, *Anal. Quant. Cytol. Histol.* **4**(2) (1982), 81–89.

[3] S. Birch, G. Stoddart and F. Beland, Modelling the community as a determinant of health, *Can. J. Public Health* **89**(6) (1998), 402–405.

[4] P. Burton, L. Gurrin and P. Sly, Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling, *Stat. Med.* **17**(11) (1998), 1261–1291.

[5] E.L. Cusick, S.W. Ewen, Z.H. Krukowski and N.A. Matheson, DNA aneuploidy in follicular thyroid neoplasia, *Br. J. Surg.* **78**(1) (1991), 94–96.

[6] J. de Leeuw, I.G.G. Kreft, Software for multilevel analysis, in: *Multilevel Modelling of Health Statistics,* A.H. Leyland and H. Goldstein, eds, Wiley, Chichester, 2001, pp. 187–204.

[7] A.V. Diez Roux, Bringing context back into epidemiology: variables and fallacies in multilevel analysis, *Am. J. Public Health* **88**(2) (1998), 216–222.

[8] A.V. Diez Roux, B.G. Link and M.E. Northridge, A multilevel analysis of income inequality and cardiovascular disease risk factors, *Soc. Sci. Med.* **50**(5) (2000), 673–687.

[9] T.E. Duncan, S.C. Duncan and H. Hops, Latent variable modeling of longitudinal and multilevel alcohol use data, *J. Stud. Alcohol* **59**(4) (1998), 399–408.

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

[11] W. Feller, *Introduction to Probability Theory and Its Applications*, Wiley, New York, 1991, p. 258.

[12] H. Goldstein, *Multilevel Statistical Models*, Internet ed., Edward Arnold, London, 1999.

[13] H.J. Gundersen and R. Osterby, Optimizing sampling efficiency of stereological studies in biology: or 'do more less well!', *J. Microsc.* **121** (1981), 65–73.

[14] H.J. Gundersen and R. Osterby, Sampling efficiency and biological variation in stereology, *Mikroskopie* **37**(Suppl.) (1980), 143–148.

[15] M. Gupta, T.M. Mayhew, K.S. Bedi, A.K. Sharma and F.H. White, Inter-animal variation and its influence on the overall precision of morphometric estimates based on nested sampling designs, *J. Microsc.* **131** (1983), 147–154.

[16] G. Haroske, J.P.A. Baak, H. Danielsen, F. Giroud, A. Gschwendtner, M. Oberholzer, A. Reith, P. Spieler and A. Böcking, Fourth updated ESACP consensus report on diagnostic DNA image cytometry, *Analyt. Cell. Pathol.* **23**(2) (2001), 89–96.

[17] C. Hedinger, E.D. Williams and L.H. Sobin, *Histological Typing of Thyroid Tumours*, Springer, Berlin, 1988.

[18] J.J. Hox, *Applied Multilevel Analysis*, TT-Publikaties, Amsterdam, 1995.

[19] C.J.M. Maas and J.J. Hox, Sample sizes for multilevel modeling, in: *Social Science Methodology in the New Millennium. Proceedings of the Fifth International Conference on Logic and Methodology. Second Expanded Edition*, J. Blasius, J. Hox, E. de Leeuw and P. Schmidt, eds, Leske + Budrich Verlag, Opladen, RG, 2002 (CD-ROM). Paper available at http://www.fss.uu.nl/ms/jh/publist/simnorm1.pdf.

[20] C. MacAulay and B. Palcic, Fractal texture features based on optical density surface area, *Anal. Quant. Cytol. Histol.* **12**(6) (1990), 394–398.

[21] D.W. Matteson, J.A. Burr and J.R. Marshall, Infant mortality: a multi-level analysis of individual and community risk factors, *Soc. Sci. Med.* **47**(11) (1998), 1841–1854.

[22] E.C.M. Mommers, N. Poulin, C.J.L.M. Meijer, J.P.A. Baak and P.J. van Diest, Malignancy-associated changes in breast tissue detected by image cytometry, *Analyt. Cell. Pathol.* **20**(4) (2000), 187–196.

[23] B. Nadjari, H. Motherby, T. Pooschke, S. Pooschke, H.E. Gabbert, D. Simon, H.D. Roher, J. Feldkamp, L. Tharandt and A. Bocking, DNA aneuploidy as a specific marker of neoplastic cells in FNAB of the thyroid, *Anal. Quant. Cytol. Histol.* **21**(6) (1999), 481–488.

[24] B. Nielsen, F. Albregtsen, W. Kildal and H.E. Danielsen, Prognostic classification of early ovarian cancer based on very low dimensionality adaptive texture feature vectors from cell nuclei from monolayers and histological sections, *Analyt. Cell. Pathol.* **23**(2) (2001), 75–89.

[25] R.Z. Omar, E.M. Wright, R.M. Turner and S.G. Thompson, Analysing repeated measurements data: a practical comparison of methods, *Stat. Med.* **18**(13) (1999), 1587–1603.

[26] T. Oyama, A.L. Vickery, Jr, F.I. Preffer and R.B. Colvin, A comparative study of flow cytometry and histopathologic findings in thyroid follicular carcinomas and adenomas, *Hum. Pathol.* **25**(3) (1994), 271–275.

[27] J. Rabash, W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford and T. Lewis, *A User's Guide to MLwiN*, Univ. of London, London, 2000.

[28] J. Shay, Economy of effort in electron microscope morphometry, *Am. J. Pathol.* **81** (1975), 503–511.

[29] A.B. Snijders and R.J. Bosker, *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publishers, London, 1999.

[30] SPSS Inc., Linear mixed-effect modeling in SPSS, An introduction to the MIXED procedure, Technical report, 2002, available at http://www.spss.com/home_page/wp127.htm.

[31] A. Stuart, *The Ideas of Sampling*, Macmillan Publishing Company, New York, 1984.

[32] L.M. Sullivan, K.A. Dukes and E. Losina, Tutorial in biostatistics. An introduction to hierarchical linear modelling, *Stat. Med.* **18**(7) (1999), 855–888.

[33] O. Tsybrovskyy and A. Berghold, Primary unit for statistical analysis in morphometry: patient or cell?, *Analyt. Cell. Pathol.* **18**(4) (1999), 191–202.

[34] A.M. van Driel Kulker, W.E. Mesker, M.J. van der Burg and J.S. Ploem, Preparation of cells from paraffin-embedded tissue for cytometry and cytomorphologic evaluation, *Anal. Quant. Cytol. Histol.* **9**(3) (1987), 225–231.

[35] N. Wang, C. Wilkin, A. Böcking and B. Tribukait, Evaluation of tumor heterogeneity of prostate carcinoma by flow- and image DNA cytometry and histopathological grading, *Analyt. Cell. Pathol.* **20**(1) (2000), 49–63.

[36] J.S. Witte, S. Greenland and L.L. Kim, Software for hierarchical modeling of epidemiologic data, *Epidemiology* **9**(5) (1998), 563–566.

[37] M. Yang, A review of random effects modelling in SAS (release 8.2), (2003), Paper available at http://multilevel.ioe.ac.uk/softrev/reviewsas.pdf.

[38] J. Zedenius, G. Auer, M. Backdahl, U. Falkmer, L. Grimelius, G. Lundell and G. Wallin, Follicular tumors of the thyroid gland: diagnosis, clinical aspects and nuclear DNA analysis, *World J. Surg.* **16**(4) (1992), 589–594.

[39] X.H. Zhou, A.J. Perkins and S.L. Hiu, Comparisons of software for generalized linear multilevel models, *The American Statistician* **53**(3) (1999), 282–290.