ACTA STEREOL 1983; 2: 298-306
STEREOLOGY AND MORPHOMETRY IN PATHOLOGY
Proc 2nd Symp Morphometry in Morphol Diagnosis
Collan Y. et al., eds.
Kuopio University Press, Kuopio 1983

REPRODUCIBILITY AND SOURCES OF VARIATION IN DIAGNOSTIC HISTOPATHOLOGY AND IN DIAGNOSTIC MORPHOMETRY

Yrjö Collan

Department of Pathology, University of Kuopio,
P.O.B. 6, 70211 Kuopio 21, Finland

ABSTRACT

Diagnostic decisions aim at finding the entities of classification which most intimately correspond to the diseases the patients are suffering from. Basicly diagnostic decisions do not differ from any decisions made by humans. Reproducibility of diagnostic decisions is not perfect. Morphometric measurements can help in diagnostic decisions, but one should realise that such measurements are not exempt from variation in the diagnostic situation.

TYPES OF DISEASE CLASSIFICATIONS

Diseases have been classified for hundreds of years. At first they were classified according to the symptoms they caused. Later, when more had been learned about tissue changes the diseases were classified according to the histopathological changes that were present. The pathogenesis of disease could be clarified through interpretation of these changes. In some cases the etiological agent of the disease could be isolated. The result was a classification which included entities classified according to pathogenesis or etiology, and sometimes both. This is where we are today.

Some disease entities seem to be well separated from each other. They have few histological features in common. On the other hand there are disease entities which cover a zone in a continuous spectrum of changes. The entities which cover the adjacent zones seem to be related, and at the border of two entities the differences may be minimal. However, if the entities have been defined well there is no overlap of adjacent entities.

THE DIAGNOSTIC DECISION

Diagnosis is not a black box. After thinking of it one can realise that what we usually call diagnosis is in fact a diagnostic decision. This decision is based on what we know about the patient. In histopathology we stress the findings on histological sections. Basicly our decision does not remarkably differ from a business decision.

A business decision

A company is planning its future activities. An interesting project is emerging. The director has to be sure that the company is not going to lose in the project, because there are no extra funds to cover potential losses. So he asks his advisors, all of them experts on various aspects of production or marketing: Are we going to lose in this venture?

The answers are as follows:

Mr.A.: Yes
Mr.B.: No
Mr.C.: I really cannot tell
Mr.D.: Absolutely we are going to lose
Mr.E.: We are certainly going to win

Now the director could try to decide who of these advisors is right. But he knows one thing. He could find out the right answer if he would launch the project, let time pass and look at the financial figures after the project had been completed. The second important thing he knows is that before the project there are no absolutely reliable ways of telling which answer is right or which answer is wrong.

A diagnostic decision may be very similar. Just have a look at the following example (see also Collan 1982).

A diagnostic decision

A young patient has got a subcutaneous tumor at the elbow. The surgeon removes is carefully but is not absolutely sure about that he was able to remove the tumor completely. In pathology department the sample is given to a specialist who, however, does not think he has met a tumour with corresponding histological appearance before. So he decides to ask his colleagues: Is this a malignant tumour or a benign tumor? He gets the following answers:

Dr. A.: Malignant
Dr. B.: Benign
Dr. C.: Borderline between malignant and benign
Dr. D.: Absolutely malignant
Dr. E.: Absolutely benign

Now the doctor has to decide who is right. It is difficult, but he might be helped by two things he knows are absolutely true. The first is that follow-up, without any interference from the medical community, might be able to show the nature of the tumor. On the other hand he knows that there are no absolutely reliable ways of telling the truth now. The only reliable solution is to wait and see. This could be possible only if he would decide that the tumor is benign. But if he would decide that the tumor is malignant the decision could lead to amputation or other kind of radical treatment, which could hide the real nature of the tumor. Effective treatment would completely remove a malignant tumor and also eliminate the chances for recurrencies or metastatic deposits. Thereafter the follow-up could no longer tell the difference between a malignant and a benign tumor.

There are elements corresponding to the above expert opinions in all our decisions. We think of all the alternatives and try to find the most probable one. And we do not act in a similar fashion, all of us. One director would like to start the project, another would like to look for a new one. One doctor would like to suggest that the tumor was malignant and should be treated that way, another would suggest that the tumor was benign and would need no treatment whatsoever more than what has already been done. Of course, we do not meet this kind of complicated problem every day. However, the example shows the nature of the system in which we are working.

TYPES OF VARIATION IN DIAGNOSTIC DECISIONS

From the above it should be clear that decisions made by different directors need not be similar. It should also be quite clear that different pathologists might sometimes reach different diagnostic decisions. This leads to variation in diagnostic decisions. Basicly we can speak of two main types of variation. One is lack of reproducibility - a performance of one pathologist is not necessarily perfectly reproduced by other pathologists or by the same pathologist if he analyses a sample he already analysed a time ago. The other type of imperfection is that caused by bias. The result is accurate if there is no bias. In other words, accuracy is perfect if the result is the same as the perfect result. But to understand what result is perfect is often impossible. One should have an objective absolute reference. Follow-up is an absolute reference, more or less. Special stains may offer reference, especially when we want to decide the tissue from which the tumor originated. But in respect to malignancy there are few absolute ways of definitely deciding whether the performance is biased or not. On the other hand reproducibility can be measured and a diagnostic system can be developed in terms of reproducibility - sometimes also called internal consistency.

REPRODUCIBILITY IN DIAGNOSTIC DECISIONS

Many studies have been carried out on the internal consistency of diagnostic decisions (Collan 1982). In the following I will present a couple of examples demonstrating the type of variation met in the diagnostic context.

For the death certificate the basic cause of death and the immediate cause of death are recorded. The pathologist also tries to explain death by a mechanism of death, which describes how the disease started and led to death. The mechanism of death includes the above causes of death. Tarvainen and Collan (1983) studied 100 randomly selected autopsies and the diagnoses suggested in connection with these autopsies. It was possible to compare the views of the clinician and the pathologist after the autopsy, and the authors also tried to find additional medically sound alternatives for the causes of death and the mechanism of death. The mean numbers of available alternatives per one autopsy were as follows:

Basic cause of death	2.6
Immediate cause of death	1.6
Mechanism of death	3.1

These figures show that the diagnostic system is not unambiguous. Even if the diagnoses made by the doctors were the same, the way how these are combined in the death certificate results in variation, which could possibly be avoided if the rules of the game were different. So, it is not only the variation in individual decisions that counts. Also the way how these decisions are combined in the diagnostic context is important. Ringsted et al.(1978) studied grading of lesions in the uterine cervix. They had three experienced pathologists who decided about the correct diagnoses. The performance of 13 pathologists was estimated in respect to these diagnoses. Results when combined showed the probability of a pathologist giving a certain diagnosis (numbers 1-5 as column heads - these correspond to the diagnoses in line headings).

	1.	2.	3.	4.	5.
1. No changes	94	3	2	<1	<1
2. Slight atypia					
3. Severe atypia	9	12	56	21	2
4. Carcinoma in situ					
5. Carcinoma	2	<1	1	2	94

Note that the figures for slight atypia or carcinoma in situ are not shown. The table shows that the ends of the spectrum are reasonably unambiguous. However, even there only 94 per cent probability for a correct diagnosis was reached. The range of variation was wider when a sample with a diagnosis from the middle of the spectrum was chosen. When the correct diagnosis was slight atypia the probability of the diagnosis being the same after repeat experiment was 21 per cent. Carcinoma in situ reached 63 per cent.

Saxen et al. (1978) studied neoplasms of the thyroid gland. Saxen (1979) presented the results as follows. Four pathologists studied all samples and chose among diagnostic entities given at the left. The numbers on the right show the percentages of cases receiving 1, 2, 3, or 4 diagnostic labels from the four pathologists each of whom was allowed to make one suggestion only. The titles to the lines present the final diagnostic labels given to the tumours.

Final diagnosis	Number of diagnoses given 1	2	3	4
Follicular carcinoma	26.1	58.5	13.5	1.8
Papillary carcinoma	81.0	15.3	2.8	0.9
Anaplastic carcinoma	50.4	38.2	7.3	4.1
Medullary carcinoma	45.2	32.3	19.4	3.2
Sarcoma	14.3	85.7	-	-
Not thyroid cancer	28.7	52.9	12.6	5.7

The above results show that certain tumor types are subject to larger diagnostic variation than others. It is also evident that classifications which try to cover all tumor types of a specific organ need not be very successful in doing that. Also this model example of diagnostic decisionmaking contains one variation element from the classification scheme, another from the subjective interpretation.

Estimation of reproducibility in diagnostic histopathology is also possible by applying the kappa statistic (Selkäinaho 1983, Holman et al. 1983, Silcocks 1983).

VARIATION IN DIAGNOSTIC MORPHOMETRY

It is like a tradition that morphometric research relies on the relative standard error (RSE) of the measurements as the sole guide to sample size. Most researchers doing stereologic or morphometric work are blinded by the more or less complicated statistics linked with the evaluation of RSE. They do not see other possible criteria. Such criteria are related to why and for what purpose we would like to make the investigation. This point needs consideration because if we rely on RSE as the sole guide to sample size, the methods often turn out to be too laborious for the practical context.

First of all we could say that variation in the absolute parameters we are trying to evaluate should rather be larger than smaller than the inherent variation in our measuring system. This approach sounds sensible because background variation in measurements may hinder us from detecting changes without extremely laborious measurements. If the variation range of the parameters in absolute units (in contrast to measured units) is small then a measuring system is needed in which the variation range is also small. It may sometimes be extremely difficult to detect any changes whatsoever between control and experimental groups. If the system of measurement is then refined and the inherent RSE of the measurements made as small as is possible under the circumstances the chances increase that changes are detected after all (especially if and when they are present). On the other hand, if the variation range of absolute parameters is huge, the measuring system can be crude and still detect changes (Selkäinaho and Collan 1983), even though individual measurements are less accurate than in the former case. An example of the former case would be an investigation in which it would be necessary to detect changes of about 10% in the thickness of the basement membrane of the glomeruli. An example of the latter case could be an increase of 10 times in the volume fraction of the epithelium - a difference that could be possible between some benign and malignant tumours.

In diagnostic morphometry if we deal with completely different morphologic entities, which are not parts of a continuous spectrum, our approach may lead to a distinction between two entities after measurement of a single parameter. On the other hand if we deal with grades as parts or zones of a continuous spectrum of changes, it is also theoretically impossible to reach an absolute distinction with a single measurement only. In the latter cases distinction may be possible when several parameters are considered simultaneously (Bezemer, Baak, deWith 1977).

A more general approach to the validity evaluation of a test system than determination of RSE , is statistical estimation of reproducibility. Reproducibility of morphometric measurements, in which also human factors may be involved, can be estimated with intraclass correlation coefficient, ICC. If the results of measurements are condensed into grades, kappa statistic can be applied (Selkäinaho 1983). These methods have been described in detail in original research papers (Cochran 1969, Selkäinaho 1983) and are also described in the presentation of Dr. Selkäinaho in this symposium (Selkäinaho 1983). In fact our approach can be more or less empirical in that we can take a representative group of samples for analysis and decide about the needs of our test system in respect to the purpose of our study. Then we can make morphometric measurements with several numbers of probes (points, unit length lines etc.). At the results we can then decide which number of probes should be taken for the test proper. This approach is good because the final needs of our test system can be made to influence the selection of the method. The approach corresponds the approach which aims at determination of the range of total variation in a morphometric measurement system -- also an alternative to RSE-based calculations. No doubt efforts can be saved with this approach, especially when we deal with a test situation in which distinction of non-overlapping entities is considered.

EXAMPLES OF VARIATION IN MORPHOMETRIC HISTOPATHOLOGY

Our research group has studied the various sources of variation in diagnostic histopathology and morphometry. These studies have dealt with various aspects of morphometric histopathology (Collan et al. 1982, Kosma et al. 1983, Kosma et al. 1983, Romppanen et al.1982). I will here give a short introduction to the subject by showing you some data on kidney biopsy analysis. We have used a test system for diagnostic purposes (Romppanen and Collan 1981, Laakso et al. 1982). Several pathologists have measured kidney biopsy samples and we have then determined the variation (expressed as coefficient of variation or CV, CV = SD/mean) of the measurements. Interobserver variation (left column) and intraobserver variation (right column) has been determined separately.

Results for three parameters are as follows:

	Coefficient of variation	
A. No. of nuclei per area	11.0 - 18.9	2.0 - 9.0
B. Surface density of GBM	8.3 - 27.1	1.9 - 8.1
C. Volume fraction of mesangium	14.8 - 25.8	6.2 - 15.0

These results have surprised some researchers. However, this is exactly what one gets when morphometry is applied in the diagnostic context. Diagnostic morphometry is full of variation sources, as is diagnostic histopathology in general. There is a tendency to consider morphometry part of exact sciences. In practical context this is not possible. Surprisingly low values of variation ranges are given in research papers on morphometry. Such variation ranges cannot be applied to the diagnostic situation but only to parts of it. In fact it is not only the complexity of the diagnostic situation that is usually forgotten but also the nature of the practical diagnostic situation. I cannot go into details here more than to say that in medicine we should not speak of retrospective and prospective studies only but also of diagnostic studies. The latter studies have the largest variation - and the greatest relevance to the patient (Collan 1983).

When we take the diagnostic approach and want to make morphometry help us we necessarily have to consider the sources of the variation. One of these is the thickness of the section. It varies, but if we know the limits within which it varies we can determine the limits within which the variation of our test system operates. The following list gives the data of the parameters described in the former list. Each parameter has two lines - the upper line gives the mean absolute values and the lower line gives the coefficient of variation:

	Thickness of sections (um)				
	1	2	3	4	5
A.	6334	6572	7724	8518	8936
	1.9	1.9	3.4	5.0	8.0
B.	0.576	0.601	0.533	0.411	0.272
	4.9	6.8	2.5	1.9	11.0
C.	11.1	11.7	13.7	13.5	12.7
	2.3	2.1	3.1	6.3	8.6

It is not difficult to see that morphometric parameters change when section thickness changes. Also the variation between measurements (here measured as CV) changes with section thickness. After learning the limits within which the variation takes place this source of variation can be handled in figures. A corresponding approach can be taken to study sample size in kidney biopsy (Romppanen et al. 1982).

REFERENCES

Bezemer PD, Baak JPA, deWith C: Discriminant analysis, exemplified with quantitative features of the endometrium. Eur J Obstet Gynec Reprod Biol 1977; 7/3: 209 - 214

Cochran WG: Errors of measurements in statistics. Technometrics 1969; 10: 637 - 666

Collan Y: Reproducibility, the neglected cornerstone of medical diagnostics. In: Collan Y, Romppanen T, eds. Morphometry in Morphological Diagnosis. Pp. 5 - 21. Kuopio University Press, Kuopio 1982

Collan Y: Morphometry in pathology: Another look at diagnostic histopathology. 9th Eur Congr Pathology, Hamburg 1983

Collan Y, Romppanen T, Karhunen J, Jantunen E: Effects of section thickness on morphometrical analysis of kidney biopsies. International Society for Stereology Meeting, Sheffield 1982

Holman CDJ, Matz LR, Finlay-Jones LR, Waters ED, Blackwell JB, Joyce PR, Kelsall GRH, Shilkin KB, Cullity GJ, Williams KE, Matthews MLV, Armstrong BK: Inter-observer variation in the histopathological reporting of Hodgkin's disease: an analysis of diagnostic subcomponents using kappa statistics. Histopathology 1983; 7: 399 - 407

Kosma VM, Collan Y, Aalto ML, Seppä A, Rautiainen M, Selkäinaho K: Reproducibility and variation in morphometric assessment of positive staining for CEA in ovarian tumours. 6th Int Congr Stereol, Abstracts p. 21. Gainesville, Fla. 1983

Kosma VM, Collan Y, Syrjänen K, Aalto ML, Seppä A, Selkäinaho K: Observer variation and reproducibility of grading: analysis of the postcapillary venules in human axillary lymph nodes using subjective and morphometric methods. Acta Stereol 1983; 2: 342 - 348

Laakso M, Pentikäinen PJ, Lampainen E, Romppanen T, Naukkarinen A, Collan Y: Trauma, renal vein thrombosis and subsequent nephrotic syndrome: a case report. Ann Clin Res 1982; 90: 20 - 28

Ringsted J, Amtrup F, Asklund C, Baunsgaard P, Christensen HE, Hansen L, Jakobsen C, Jensen NK, Moesner J, Rasmussen J, Reintoft I, Rolschau J, Starklint H, Thommesen N, Vrang J: Reliability of histo-pathological diagnosis of squamous epithelial changes of the uterine cervix. Acta Pathol Microbiol Scand A 1978; 86: 273 - 278

Romppanen T, Collan Y: Morphometrical method for analysis of kidney biopsies in diagnostic histopathology. Proc 3rd Eur Symp Stereol, Ljubljana. Stereol Iugosl 1981; Suppl. 1: 435 - 442

Romppanen T, Karhunen J, Jantunen E, Collan Y: Interpretation of kidney biopsy: Representative number of glomeruli for evaluation of various glomerular parameters. International Society for Stereology Meeting Sheffield 1982

Saxen E: Histopathology in cancer epidemiology. The Maude Abbott Lecture. Pathology Annual 1979; 203 - 217

Saxen E, Franssila K, Bjarnason O, Normann T, Ringertz N: Observer variation in histologic classification of thyroid cancer. Acta Pathol Microbiol Scand A 1978; 86: 483 - 486

Selkäinaho K: Deriving coefficients of internal consistency of measurements: ICC and kappa. Reports on Statistics, University of Jyväskylä 1983; 12: 1 - 16

Selkäinaho K: Statistics in stereology and morphometry. Acta Stereol 1983; 2: 239 - 249

Selkäinaho K, Collan Y: Reproducibility of stereologic and morphometric measurements. Statistical considerations. 6th Int Congr Stereol, Abstracts p.5. Gainesville, Fla. 1983

Silcocks PBS: Measuring repeatability and validity of histological diagnosis - a brief review with some practical examples. J Clin Pathol 1983; 36: 1269 - 1275

Tarvainen I, Collan Y: Kuolemansyyn määrittämisen ongelma. Cause of death - decision problematics. Suomen Lääkärilehti 1983; 38: 1807 - 1811