Virchows Arch (2000) 436:229–233

ORIGINAL ARTICLE

Claus Fenger · Morten Frisch · Jeremy J. Jass
Geraint T. Williams · Jørgen Hilden

Anal cancer subtype reproducibility study

Received: 9 July 1999 / Accepted: 6 October 1999

Abstract For histological subtyping of anal squamous carcinomas the WHO advocates a six-way subdivision, but it has been suspected that the six types cannot be reliably discriminated in practice. We conducted a blinded study involving slides from 103 consecutive cases, each slide being examined by three experts (from Denmark, Australia and UK) on two occasions at least 8 months apart. Agreement on subtypes was low: 72% between rounds within pathologist, 61% between pathologists. Even for the commonest, and most stably diagnosed, type, viz. large-cell keratinising squamous carcinoma, the intra- and interpathologist frequencies of confirmation were only 81% and 71%, respectively. The pathologist marked the picture as typical and his subtype diagnosis as certain 41% of times: even then confirmation frequencies were only 88% and 74%, respectively. Calculations, including kappa analyses, suggest that 26% of the typing variation was noise. The WHO scheme must be even more unreliable in everyday practice. We finally mention a recently demonstrated link between human papilloma virus (HPV) and certain types of anal cancer, which may well provide an additional argument for revising existing subtyping schemes.

Key words Anus · Carcinoma · Histology · Observer variation · Squamous cell carcinoma

Parts of this study were presented at the XVth European Congress of Pathology, Copenhagen 1995

C. Fenger (✉)
Department of Pathology, Odense University Hospital,
DK-5000 Odense C, Denmark
e-mail: claus.fenger@ouh.dk
Tel.:+45-6541-4807, Fax: +45-6591-2943

M. Frisch
Department of Epidemiology Research,
Danish Epidemiology Science Centre, Statens Seruminstitut,
Copenhagen, Denmark

J.J. Jass
Department of Pathology, University of Queensland Medical School,
Herston Qld, Australia

G.T. Williams
Department of Pathology,
University of Wales College of Medicine, Cardiff, UK

J. Hilden
Department of Biostatistics, Panuminstituttet,
Faculty of Health Sciences, University of Copenhagen, Denmark

Introduction

The anal canal extends from the upper to the lower border of the internal anal sphincter, i.e. from the pelvic floor to the anal opening. The mucosa can be divided into three zones according to the epithelial lining. The upper zone is covered by colorectal type mucosa and the lower zone with unkeratinised squamous epithelium, which gradually merges into the perianal skin. The middle zone is called the anal transitional zone (ATZ) and extends from the dentate line (DL) approximately 1 cm upwards [4]. Tumours may arise in all three zones and in the perianal skin.

Malignant epithelial tumours of the anal canal and anus comprise squamous variants, adenocarcinoma variants and malignant melanoma. According to the WHO [7], the squamous variants can be divided into five different types and the usual group of “others”. The distinction between these types is not always clear, and the reported relative incidences have shown large differences. Thus, the proportion of basaloid carcinoma has varied from 10% to almost 70% [2, 3], and morphological features considered typical for the different variants have often been present in the same tumours [15].

In the present study we have tested the intraobserver and interobserver variation in typing anal carcinoma of the squamous type, between three histopathologists with a particular interest in gastrointestinal pathology.

Materials and methods

Tissue specimens

Paraffin-embedded, HE-stained tissue specimens were retrieved for 103 patients with anal squamous carcinomas (68 women, 35

men) consecutively diagnosed in 15 Danish pathology departments during the years 1988–1992. To obtain sufficient amounts of material for comprehensive histological evaluation, we included tumour tissue only from tumours removed by means of abdominoperineal resection (anorectal amputation, *n*=51), or from local excisions, including large (>5 mm on the smallest side) surgical biopsies (*n*=52).

Procedure

A single slide from each of the 103 cases was selected for the study. They had original labels removed to make them as unrecognisable as possible. Subsequently, slides were numbered in random order and examined by each of three pathologists (C.F., J.J.J. and G.T.W.). After the first round, slides were renumbered by one of us (M.F.), thus ensuring that no observer could possibly know about his histological evaluation from the first round. To further minimise the potential for recall between rounds, each observer performed the second histological evaluation of the 103 specimens at least 8 months after his first examination.

Histological evaluation

The observers classified each specimen according to a six-way distinction between subtypes of anal squamous carcinomas recommended by the WHO [7], which comprises

1. Large-cell keratinising squamous carcinoma (LCKSC)
2. Large-cell nonkeratinising squamous carcinoma (LCnKSC)
3. Basaloid carcinoma (BC)
4. Verrucous carcinoma (VC)
5. Basal cell carcinoma of skin (BCC)
6. Other types (OT)

Moreover, for each specimen, observers were asked to state whether they considered the specimen to be (1) typical, (2) somewhat typical or (3) not typical of the histological category chosen (typicality assessment). Finally, observers should note whether they felt their choice of histological category was (1) certain, (2) tentative or (3) uncertain (certainty assessment).

Statistics

In the absence of a histological 'gold standard', disagreement is the main issue. From a practical point of view, what one would like to know is how often the subtype diagnosis made by one pathologist is confirmed or contradicted by another; and, in the case of a contradiction, with what frequency the second observer will 'reallocate' the specimen to this or that alternative subtype. In particular, one would like to know when the chances of confirmation are good or poor, which amounts to asking: how do the frequencies of confirmation or reallocation vary with the first observer's answer (the diagnosis he/she made and the typicality and certainty scores he/she marked)? While *inter*pathologist disagreement is what matters in practice, it is put in perspective by the extent to which pathologists fail to confirm their own diagnoses (*intra*pathologist disagreement).

Further insight is gained by a kappa-type analysis. Part of interpathologist disagreement is marginal disagreement, by which is meant that the observers do not use the subtype labels equally often: in statistical terms, their marginal answer counts are said to differ. Agreement on every specimen is then unachievable. The maximum agreement compatible with the marginal counts (abbreviated to max. below) reflects the extent to which the margins agree. (To illustrate this point, suppose two cardiologists were asked to classify 100 ECGs as normal or abnormal. If we are told that one of them rated 40, and the other 55, as abnormal we know that, at best, they could be agreeing on abnormality in 40 and on normality in 45 cases: inspection of their paired answers would reveal an 85% agreement, at the most. By fiddling with numbers one also finds that they must be agreeing in at least 5 cases, so max. and min., which are the maximum and minimum agreement compatible with the marginal counts, become: max. = 85%, min. = 5%.)

For proper appreciation of an observed frequency of agreement (obs. below) one needs three reference quantities which characterise the two marginal distributions: max. and min. have just been explained; and cex is the chance-explainable agreement that would arise – "by pure chance" – by random pairing of the diagnostic answers. Given the marginal counts, this is the hypothetical level of agreement that would be expected if the pathologists did not extract any shared typing clues from the slides. Normally obs>cex, so we have:

$$0 \le \min < \text{cex} < \text{obs} \le \max \le 1$$

The kappa value is defined as (obs–cex)/(1–cex), thus representing the fraction of the (hypothetical) pure-chance disagreement that the pathologist pair has managed to avoid. The reference quantities have been similarly rescaled.

Intra-pathologist kappa analyses were based on the marginal distributions of the two rounds of microscopy. Averaged kappa values were calculated from component kappas as the sum of the numerators divided by the sum of the denominators.

Presentation

We have not given *P*-values or confidence ranges, as the message of the data is clear enough without them. However, all patterns mentioned in the Results section are highly significant. As to the ordering of the six histological subtypes in the tables, we adopted the one shown in Table 1 because, as we foresaw, most of the diagnostic hesitation then concerns adjacent subtypes.

Results

Table 1 shows the distribution of the histological diagnoses. The label BCC was rarely used. Pathologist P3 uses LCnKSC rather infrequently, presumably opting for the neighbouring subtypes when there is no clear keratinisation. In the second round (R2), the same applies to P1 and P2, too, the frequency of LCnKSC thereby dropping appreciably. The analyses that follow must treat this feature as a reflection of fortuitous fluctuations in diagnostic style, but actually a systematic change of behaviour may well be suspected.

Disagreement is frequent, but, as shown in Table 2, most disagreement is confined to adjacent subtypes; the straightforward calculations involved are explained in the note. Understandably, the miscellaneous category, OT, is a partial exception. Overall, the intrapathologist frequency of disagreement is 28%, and disagreement increased to 39% when another pathologist examined the slides. Even with large-cell keratinising squamous carcinoma, which appears to be the most stably diagnosed condition according to Table 2, the corresponding rates of nonconfirmation are 19% and 29%.

As might be expected, the typicality and certainty scores were highly correlated, mutually as well as with subtype agreement (data not shown). It turned out that on 41% of occasions (253 microscopies out of 103×3×2=618) the pathologist both found the specimen typical and felt certain of his diagnosis, and it would be

Table 1 Overview of classifications: 103 anal carcinoma specimens were classified by three pathologists, *P1, P2, P3*, in two rounds, *R1, R2* (*VC* verrucous carcinoma, *LC(n)KSC* large-cell (non)keratinising carcinoma, *BC* basaloid carcinoma, *OT* other types, *BCC* basal cell carcinoma of skin)

	Histological diagnosis						Total
	VC	LCKSC	LCnKSC	BC	OT	BCC	
Pathologist							
P1	1	73	66	61	5	–	2*n*
P2	18	56	58	61	11	2	2*n*
P3	17	76	19	87	5	2	2*n*
Round							
R1	20	95	95	90	8	1	3*n*
R2	16	110	48	119	13	3	3*n*
Total	36	205	143	209	21	4	6*n*
(%)	6%	33%	23%	34%	3%	1%	100%

Table 2 Confirmation vs reallocation (intrapathologist and, *after semicolon,* interpathologist). The percentages (*boldface*) along the diagonal are confirmation frequencies: they show that, regardless of a pathologist's diagnosis, the estimated chance of obtaining the same answer is never more than 81% with the same pathologist (in the other round) and never more than 71% with another pathologist. Most disagreement involves neighbouring subtypes

	Estimated percentage probabilities of the label on the left given the current allocation (column label)					
	VC	LCKSC	LCnKSC	BC	OT	(BCC[b])
VC	***72;28***	4;10	1;2	0;0	0;0	(0;0)
LCKSC	25;59	***81;71***	14;16	4;6	10;11	(25;6)
LCnKSC	3;10	10;11[a]	***60;49***	16;21	10;10	(0;19)
BC	0;3	4;6	23;30	***77;68***	38;32	(0;38)
OT	0;0	1;1	1;1	4;3	***38;45***	(25;12)
BCC	0;0	0;0	1;1	0;1	5;2	***(50;25)***
Total	100;100	100;100	100;100	100;100	100;100	100;100

[a] Calculation: on 205 occasions a pathologist assigned a specimen to LCKSC (Table 1). When seen in his other round, 20 of the specimens concerned were assigned to LCnKSC (10%); the other pathologists saw those specimens on 2×2×205=820 occasions and chose LCnKSC 91 times (11%)

[b] Frequencies based on very small numbers

Table 3 Actual percentage agreement (*Obs*), with reference quantities and kappa equivalents. The values obtained for a specific pathologist, or pair of pathologists, were similar, and therefore only the averaged values are given. (*Cex* "chance-expectable" level of agreement, i.e., the one that is used as baseline in the kappa calculation; cf. Methods section)

	Actual				Complete Agreement
	Min.[a]	Cex	Obs	Max.	
Between the rounds of a given pathologist					
Agreement (%)	0	29	**72**	83	100
Kappa equivalents	–0.41	0	**0.61**	0.76	1
Between pairs of pathologists					
Agreement (%)	0	27	**61**	81	100
Kappa equivalents	–0.38	0	**0.47**	0.74	1

[a] Complete disagreement and Min. coincide, reflecting the fact that, in this data set, the marginals do not preclude 100% disagreement

reassuring to learn that, in this situation at least, there is little risk of subtype disagreement. Even here, however, the pathologist was contradicted 12% of the times by himself, and no less than 26% of the times when a colleague saw the slide; the calculation is analogous to the one explained in Table 2.

Table 3 shows that performance is equally disappointing when expressed in terms of kappa. The intrapathologist average agreement is 72%, but of the 28% disagreement, 17% (=100–83%) are due to between-rounds changes in the marginal distribution of the subtypes. Marginal disagreement also accounts for about one half of the inter-pathologist disagreement since the max. value (81%) is located midway between obs. = 61% and 100%.

A more sophisticated analysis, which takes the pairwise discriminabilities of the histological subtypes into account, suggests that at least 26% of the response variation is variation peculiar to the observer or the situation.

In other words, when one compares the pathologists' job to listening to a noisy radio channel, it is as is for every 74 units of diagnostic signal elicited from the specimens, at least 26 units of noise was injected into the process.

Discussion

Typing of tumours is particularly relevant if the individual types show different biological course or require a different treatment. Typing is also of significance if special aetiological factors play a role in one or more but not all types. Finally, typing is natural when there are clear histological differences between tumours, as this implies a difference in the structure or function of the genomes of the tumour cells and therefore could indicate a different histiogenesis or a biological significance not yet proven.

In anal carcinomas the prognosis is consistently reported to be related to tumour size, depth of spread and node involvement [11], but only exceptionally to the commonly described subtypes [9]. It is, however, generally accepted that the prognosis is poorer for the rare mucoepidermoid carcinoma or squamous carcinoma with mucinous microcysts and the small cell anaplastic carcinoma [11].

At the time of diagnosis most anal carcinomas have grown to a size that disguises their point of origin, and the distinction between anal canal and anal margin tumour is therefore often impossible, as is a distinction between tumours arising in the ATZ or squamous epithelium. The treatment is now radiotherapy in combination with chemotherapy, and only small tumours located entirely below the dentate line can be treated with local excision. The histopathological diagnosis and typing is therefore now almost entirely based on biopsies. A reproducibility study on biopsy material could show a better agreement than that obtained on surgical specimens, because the smaller areas investigated would probably show less histological variation. On the other hand, this would most probably lead to a less correct typing of the tumours as a whole.

Anal carcinomas of the squamous variants have long been suspected to be linked with sexually transmitted disease, and associations with lymphogranuloma inguinale, syphilis, gonorrhoea, herpes simplex virus type 2 and *Chlamydia trachomatis* have been reported [3]. A recent large-scale case-control study in Denmark and Sweden yielded substantial evidence that a sexually transmitted infection is a crucial risk factor for the development of anal cancer. By means of the polymerase chain reaction, human papillomavirus (HPV) DNA was detected in the vast majority of anal cancers, but in none of a control series of rectal adenocarcinomas [6].

The material just mentioned has also provided evidence that anal carcinomas can be divided into two main groups according to their location and possible origin. Thus, tumours in the anal canal are characterised by relatively small cells, basaloid features and lack of keratinisation and, in almost all cases, show a positive reaction for oncogenic types of HPV, with HPV 16 as the most prevalent type. In contrast, most perianal tumours are of the large-cell keratinising type and only show positivity for HPV in about two thirds of the cases [5]. This is in accordance with observations in earlier studies based on smaller series [10, 14].

This recent insight was not available when the present study was planned, let alone when the WHO subtypes were defined, and our data cannot be used to answer the twin questions it raises, namely (1) whether, if a clear demarcation turns out to exist between HPV and non-HPV carcinomas, clinical histopathologists will be able to discriminate between them reproducibly in practice; and (2) whether microscopy is sufficiently reliable to help us in finding out whether or not such a demarcation exists.

Anyhow, neither basaloid carcinomas, most of which contain DNA from oncogenic HPV types, nor large-cell keratinising squamous cell carcinomas, which have a much looser link to HPV [5], are particularly reproducible. Both blend into the intervening third major subtype, the large-cell *non*keratinising squamous cell carcinomas (Tables 1, 2). However, discrimination between the keratinising subtypes (LCKSC and VC) on the one side and basaloid carcinoma (BC) on the other is perhaps acceptable.

Reproducibility of histopathological diagnoses is crucial for their utility in daily practice as well as in science. In the last decade numerous reports have appeared describing tests of various classification and grading systems and giving results obtained by means of kappa statistics [12]. Such studies have shown kappa values ranging from close to zero up to more than 0.80. At least two studies have dealt with WHO tumour typing within four to six categories and have shown kappa values ranging from 0.39 for the typing of prostate carcinoma [13] to 0.62–0.87 for the typing of endometrial carcinoma [8]. Grading of anal intraepithelial neoplasia is subject to a similar uncertainty [1].

Kappa values are correlation-like coefficients of agreement and may be viewed as abstract correlation coefficients. To be specific, our interobserver value of 0.47 indicates that the 'observer–truth correlation' is at most $\sqrt{0.47}=0.68$, which in turn implies that 32% of the response variation is effectively noise.

Admittedly, the kappa coefficients in Table 3 are based on all six subtypes being equally distant from one another, as if all types of typing disagreement were equally serious medically. In theory, unequal pairwise distances between subtypes could be introduced ('weighted kappa values'). Such distances should ideally reflect the seriousness of typing mistakes, be it for therapeutic purposes or for prognostic counselling. We have not imposed a distance scheme on the current six-way WHO classification system for anal cancer, but calculations have made it clear what would be the result of doing so: whatever distance scheme one were to adopt, a sizable fraction, at least 26%, of the response variation would be noise.

In clinical chemistry one would *not* accept such low correlations between the measurement and the true concentration. Although the interpretation of kappa is always

dubious, and although there is no way of deciding how large kappa ought to be, save via a detailed analysis of the clinical consequences of typing disagreement, it can probably be said that, with a dedicated international team of anal cancer experts such as those involved in the present study, kappa should exceed 0.9. The six-category WHO system did not attain this level of reproducibility. Also, it may not fit in with the aetiological evidence that is beginning to emerge. Eventually this may mean that it will lose even the limited therapeutic relevance it has been believed to have. In a few years it will probably have to be revised.

References

1. Carter PS, Sheffield JP, Shepherd N, et al (1994) Interobserver variation in the reporting of the histopathological grading of anal intraepithelial neoplasia. J Clin Pathol 47:1032–1034
2. Deans GT, McAleer JJA, Spence RAJ (1994) Malignant anal tumours. Br J Surg 81:500–508
3. Fenger C (1991) Anal neoplasia and its precursors: facts and controversies. Semin Diagn Pathol 8:190–201
4. Fenger C (1997) Anal canal. In: Sternberg SS (ed) Histology for pathologists, 2nd edn. Lippincott-Raven, Philadephia, pp 551–571
5. Frisch F, Fenger C, van den Brule AJC, et al (1999) Variants of squamous cell carcinoma of the anal canal and perianal skin and their relation to human papillomaviruses. Cancer Res 59:753–757
6. Frisch M, Glimelius B, van den Brule AJC, et al (1997) Sexually transmitted infection as a cause of anal cancer. N Engl J Med 337:1350–1358
7. Jass JR, Sobin LH (1989) Histological typing of intestinal tumours, 2nd edn. Springer, Berlin Heidelberg New York, pp 41–47
8. Nedergaard L, Jacobsen M, Andersen JE (1995) Interobserver agreement for tumour type, grade of differentiation and stage in endometrial carcinomas. APMIS 103:511–518
9. Peiffert D, Bey P, Pernot M, et al (1997) Conservative treatment by irradiation of epidermoid cancers of the anal canal: prognostic factors of tumoral control and complications. Int J Radiat Oncol Biol Phys 37:313–324
10. Scholefield JH, Palmer JG, Shepherd NA, et al (1990) Clinical and pathological correlates of HPV type 16 DNA in anal cancer. Int J Colorect Dis 5:219–222
11. Shepherd NA, Scholefield JH, Love SB, et al (1990) Prognostic factors in anal squamous carcinoma: a multivariate analysis of clinical, pathological and flow cytometric parameters in 235 cases. Histopathology 16:545–555
12. Svanholm H, Starklint H, Gundersen HJG, et al (1989) Reproducibility of histomorphologic diagnosis with special reference to the kappa statistics. APMIS 97:689–698
13. Svanholm H, Starklint H, Barlebo H, Olsen S (1989) Histological evaluation of prostatic cancer. 1. Reproducibility of tumour type. APMIS 97:799–704
14. Vincent-Salomon A, de la Rochefordière A, Salmon R, et al (1996) Frequent association of human papillomavirus 16 and 18 DNA with anal squamous cell and basaloid carcinoma. Mod Pathol 9:614–620
15. Williams GR, Talbot IC (1994) Anal carcinoma: a histological review. Histopathology 25:507–516