European Commission Working Group on Breast Screening Pathology: J.P. Sloane (Chairman)
I. Amendoeira · N. Apostolikas · J.P. Bellocq
S. Bianchi · W. Boecker · G. Bussolati · D. Coleman
C.E. Connolly · V. Eusebi · C. De Miguel · P. Dervan
R. Drijkoningen · C.W. Elston · D. Faverly · A. Gad
J. Jacquemier · M. Lacerda · J. Martinez-Penuela
C. Munt · J.L. Peterse · F. Rank · M. Sylvan
V. Tsakraklides · B. Zafrani

Consistency achieved by 23 European pathologists from 12 countries in diagnosing breast disease and reporting prognostic features of carcinomas

Abstract A detailed analysis of the consistency with which pathologists from 12 different European countries diagnose and classify breast disease was undertaken as part of the quality assurance programme of the European Breast Screening Pilot Network funded by the Europe against Cancer Programme. Altogether 107 cases were examined by 23 pathologists in 4 rounds. Kappa (κ) statistics for major diagnostic categories were: benign (not otherwise specified) 0.74, atypical ductal hyperplasia (ADH) 0.27, ductal carcinoma in situ (DCIS) 0.87 and invasive carcinoma 0.94. ADH was the majority diagnosis in only 2 cases but was diagnosed by at least 2 participants in another 14, in 9 of which the majority diagnosis was benign (explaining the relatively low κ for this category), DCIS in 4 (all low nuclear
grade) and invasive carcinoma (a solitary 1-mm focus) in 1. The histological features of these cases were extremely variable; although one feature that nearly all shared was the presence of cells with small, uniform, hyperchromatic nuclei and a high nucleo-cytoplasmic ratio. The majority diagnosis was DCIS in 33 cases; \( \kappa \) for classifying by nuclear grade was 0.38 using three categories and 0.46 when only two (high and other) were used. When ADH was included with low nuclear grade DCIS there was only a slight improvement in \( \kappa \). Size measurement of DCIS was less consistent than that of invasive carcinoma. The majority diagnosis was invasive carcinoma in 57 cases, the size of the majority being 100% in 49. The remainder were either special subtypes (adenoid cystic, tubular, colloid, secretory, ductal/medullary) or possible microinvasive carcinomas. Subtyping was most consistent for mucinous (\( \kappa, 0.92 \)) and least consistent for medullary carcinomas (\( \kappa, 0.56 \)). Consistency of grading using the Nottingham method was moderate (\( \kappa=0.53 \)) and consistency of diagnosing vascular invasion, fair (\( \kappa=0.38 \)). There was no tendency for consistency to improve from one round to the next, suggesting that further improvements are unlikely without changes in guidelines or methodology.

**Key words** Breast · Invasive carcinoma · Ductal carcinoma in situ · Atypical hyperplasia · Diagnostic consistency · Kappa statistics

**Introduction**

In order to encourage breast cancer screening in the European Union, the European Commission set up a Pilot Breast Screening Network, funded under the European against Cancer Programme. Rigorous quality assurance (QA) arrangements covering all professional disciplines were put in place to ensure that the highest possible standards were reached. The European Commission Working Group on Breast Screening Pathology (ECWG BSP) was formed to deal with the pathological aspects of QA. The group produced guidelines on reporting breast specimens which were derived from those already published in the UK [4, 9] and set up a slide exchange type of external quality assessment (EQA) scheme involving, amongst others, the pathologists working in the centres funded under the pilot programme. As part of its activities the Working Group has undertaken detailed studies of the consistency with which its members diagnose breast diseases and report prognostic features.

The present study had three aims. The first was to determine whether an adequate level of consistency could be achieved by 23 pathologists from 12 European countries in diagnosing major categories of breast disease and reporting those histological features of prognostic significance that are used for determining how individual patients should be managed. This was important if results from different European screening centres were to be compared. The second was to compare the findings with those obtained several years previously in a similar study by a similar number of pathologists from the UK. This would enable us to determine whether there were any major international differences in pathological reporting among the different countries of the European Union. The third was to determine whether the considerable improvements in the guidelines used for the previous UK study had resulted in any improvement in diagnostic consistency.

**Materials and methods**

Cases were submitted to the co-ordinating centre (University of Liverpool) by the 23 members of the Working Group and were selected according to certain predetermined diagnoses made at the referring centres: benign (not otherwise specified), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS) of high, intermediate and low nuclear grades and invasive carcinoma of different types and grades. No selection based on histological appearance was made within these groups, the cases being chosen in strict chronological sequence following a specified accession date. Blocks were not used if sections of adequate quality could not be prepared from them or if the lesions they contained were of inadequate size to prepare 23 virtually identical sections without significant change in the lesion's size or histological characteristics. One H&E-stained section from each case was therefore selected by each member of the Working Group, who reported them using a pro forma and following the guidelines published by the Working Group [4]. These guidelines were derived from those produced for the UK National Health Service Breast Screening Programme and are identical to those now used in the UK [9]. The pro-forma was analysed electronically by the Cancer Screening Evaluation Unit, Sutton, Surrey, UK. The slides were not marked in any way and no specific areas were selected. A learning set of slides was not circulated at the beginning and there was no detailed discussion about how the guidelines should be followed before the study began. The cases were, however, discussed in detail after each circulation at Working Group meetings. Altogether 107 sets of H&E-stained sections were examined in four circulations.

The criteria for diagnosing ADH in the guidelines were those of Page and Rogers [11]. The classification of DCIS was based entirely on nuclear grade, as defined by Holland et al. [6], but cell polarisation was not taken into account. Invasive carcinomas were graded by the Nottingham method [3]. In situ and invasive carcinomas were measured on the circulated slides. The maximum diameter was recorded and was defined as the greatest distance between two points on the periphery of the lesion. No guidance was given on the method of performing the measurements, which was thus subject to some variation e.g. using the Vernier scale on the microscope stage, using a ruler to measure the lesion directly, with or without marking the slide.

The agreement between participants on the categorization of cases using each classification was measured by calculating \( \kappa \) statistics, which take into account the level of agreement expected purely by chance, and which also require no knowledge of the true diagnosis. For 2-way classifications there is only a single value of \( \kappa \) but for consistency of presentation this is given in all columns of the tables in this paper. For the 3-way classifications, an overall \( \kappa \) value was calculated from the \( \kappa \) values for individual categories, weighted by the proportion of reports in each category. Values of \( \kappa \) range from 0 for chance agreement only to +1 for perfect agreement, with a negative value implying systematic disagreement. Landis and Koch [8] suggest the following interpretation of different ranges of \( \kappa \): 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, 0.81–1.00 almost perfect. One disadvantage of \( \kappa \) statistics is their dependence on the prevalence of cases in each category; in particular, this will influence comparisons between different circulations.