Histological Diagnosis of Precancerous Lesions of the Stomach: A Reliability Study

MARTYN PLUMMER,* EVA BUIATTI,** GLADYS LOPEZ,† SIMON PERAZA,† JORGE VIVAS,† WALTER OLIVER† AND NUBIA MUÑOZ*

Gastric carcinogenesis is believed to be a multistage process in which the occurrence of stomach cancer is preceded by a sequence of precancerous stages: chronic gastritis, atrophy, intestinal metaplasia and dysplasia.¹ These precancerous lesions, identified by histological diagnosis of endoscopic specimens, may be used as surrogate endpoints in trials of potential chemopreventive agents, assuming that agents which positively influence progression or regression of precancerous lesions will be effective in preventing stomach cancer. The assumptions underlying the use of surrogate outcomes are discussed by Prentice.²

In ongoing trials of precancerous lesions of the stomach, the endpoint is the progression or regression rate during follow-up.³ The usefulness of this endpoint depends on the accuracy of histological diagnosis. Underlying changes in the gastric mucosa may be obscured by three sources of diagnostic misclassification which are all superimposed on each other: biopsy sampling error, observer bias and random observer error. Biopsy sampling error refers to the possibility that precancerous lesions may be missed at biopsy because of their multifocal nature. Observer bias refers to the tendency of a given pathologist to misclassify a given precancerous lesion, resulting in a consistent disagreement between two pathologists. Random observer error refers to the fact that the same pathologist reviewing the same biopsy on different occasions may give different diagnoses.

In the context of a blind randomized trial, all misclassification is random and non-differential. The net result is a loss of power to detect real effects of any chemopreventive agent on progression and regression rates.⁴ It is therefore important to assess the extent of
Materials and Methods

Subjects in the Chemoprevention Trial

The study population in the chemoprevention trial is derived from the participants in the Gastric Cancer Control Program of Tachira State, Venezuela. Eligible subjects were permanent residents of Tachira, aged 35–69 years who, after double contrast X-ray, were selected to undergo gastroscopic examination. A total of 2200 subjects have been recruited. At entry into the trial, a physical examination, blood collection, and gastroscopy were performed. During gastroscopy seven biopsies from standardized sites of the stomach were taken, two of which were frozen and five processed for histological evaluation. The study is described in detail elsewhere. The five biopsies for histological assessment were taken from the lesser curvature of the antrum, approximately 1 cm from the pylorus (biopsy A), from the greater curvature of the antrum, approximately 1 cm above the pylorus (biopsy B), from a mid-portion of the lesser curvature of the antrum (biopsy C), from the lesser curvature of the antrum immediately below the incisura (biopsy D), and from the middle corpus, approximately 2 cm from the lesser curvature (biopsy F). The two frozen biopsies (E and G) are not included in this substudy.

All gastric biopsies for histological assessment were fixed in buffered formalin and stained with haematoxylin-eosin. Those biopsies positive for intestinal metaplasia were also stained with PAS Alcian Blue and HID-Alcian Blue to determine subtypes of intestinal metaplasia. All slides were read by three pathologists who recorded their diagnosis on a standard form in terms of normal mucosa (NM), superficial gastritis (SG), chronic gastritis (CG), atrophic gastritis without intestinal metaplasia (AG) or with intestinal metaplasia (IM), and dysplasia (Dys). An overall diagnosis was also given for each subject, as the most severe among those reported. A more detailed description of the histological criteria, and the prevalence of mucosal lesions in the study population is given elsewhere. In summary, at baseline NM and SG together were represented in only 4.5% of the first 1477 participants and CG contributed 42.6%. AG without IM was represented in 20.2%, IM was reported in 26.8% and Dys in 6.0% of the subjects.

In the pilot phase of the main trial, a group of 220 subjects was involved in a randomized trial to evaluate the efficacy of Helicobacter pylori eradication. The subjects underwent gastroscopy at baseline and were then randomized to receive anti-H. pylori treatment or placebo for 2 weeks. One month after the end of treatment, the subjects returned for a follow-up visit. All subjects were offered a second gastroscopy but only 97 accepted.

Biopsies were taken according to the usual protocol and histological diagnosis was performed by the same pathologists involved in the main trial. Due to the very low eradication rate, the anti-H. pylori therapy was excluded from the main trial protocol.

Subjects in the Diagnostic Reliability Study

Two independent groups of subjects were selected to assess diagnostic reliability.

Group 1 is a stratified random sample of 50 subjects, drawn from the first 1500 subjects entering the main trial who had five complete biopsies. The sample was stratified by the original diagnosis with 10 subjects in each diagnostic category. Subjects with NM and SG were considered as a single stratum for the purpose of drawing the sample, due to the small number of subjects in this category. The biopsy results from the baseline gastroscopic examination were re-evaluated by two pathologists blind to each other and to the original diagnosis. A new diagnosis for each biopsy and a new overall diagnosis for each subject were made. Differences between the diagnoses by the two pathologists are due to a combination of observer bias and random observer error. Since both pathologists were reviewing the same slides, biopsy sampling error has been controlled in this comparison.

Group 2 consists of the 45 subjects in the placebo group of the anti-H. pylori eradication trial who underwent gastroscopy before and after treatment. The biopsies from both gastroscopic examinations were re-evaluated by pathologist X blind to the subject’s identity. It is unlikely that the status of the gastric mucosa changed in the short time interval between the two examinations, so differences between the two results are due to a combination of random observer error and biopsy sampling error. Since the same pathologist reviewed both sets of slides, observer bias has been controlled in this comparison.

Statistical Methods

In both groups, two overall diagnoses were produced for the subjects under study. The level of agreement between the two diagnostic ratings was assessed both by
using crude agreement (the proportion of subjects for whom the two diagnoses are the same) and by using Cohen’s Kappa. In addition, the two diagnoses at review in group 1 were compared with the routine baseline diagnosis.

Cohen’s Kappa measures the level of exact agreement between two categorical ratings, adjusted for the amount of agreement that would be expected if the two ratings were uncorrelated. A Kappa value of 1, the maximum, indicates perfect agreement while a value of 0 indicates only the level of agreement that would be expected from two uncorrelated rating scales. Interpretation of intermediate values is not so simple, but a set of benchmarks proposed by Landis and Koch has become a de facto standard. According to these benchmarks, the range 0–1 is divided into five categories: ‘poor’ (0–0.2), ‘fair’ (0.2–0.4), ‘moderate’ (0.4–0.6), ‘substantial’ (0.6–0.8) and ‘almost perfect’ (0.8–1). When interpreting the results, it is important to remember that Kappa is a measure of exact agreement and not of validity. Assessment of validity would require a ‘gold standard’ diagnosis, which is not available.

In addition to an assessment of agreement between overall diagnoses, agreement between the ratings for each of the five biopsies, A to D plus F was also assessed. Kappa gives a measure of agreement for the whole population which may hide differences in diagnostic reliability for different grades of precancerous lesions. To account for this, the repeat frequency was calculated for each diagnostic category. This gives the probability, conditional on the result of one diagnosis, that the other diagnosis will agree. A separate repeat frequency is calculated for each grade of precancerous lesion.

Factors which may influence agreement between the two diagnoses were assessed by using exact agreement between the two ratings as an outcome variable in a logistic regression model. The variables age, sex, original diagnosis, biopsy site and biopsy quality, as stated by the pathologist on the diagnostic form, were examined as possible predictors of agreement at review. The overall diagnosis was used first and then the results of the five individual biopsies were pooled using generalized estimating equations to account for correlation between the biopsy results of the same subject.

RESULTS

Inter-observer Agreement

In Table 1 the overall diagnoses by the two pathologists for the 50 subjects in group 1 are presented (one subject is missing). Crude agreement between the two pathologists is 76%, while Kappa is 0.69. Looking at individual biopsy sites, Kappa values for A, B, C, D and F are 0.44, 0.43, 0.57, 0.55 and 0.55, respectively. The repeat frequencies are 0.73 for SG and CG, 0.65 for AG and 0.96 for IM, while agreement is total for dysplasia.

Subjects with more advanced lesions are overrepresented in group 1, compared with the study population, due to the stratification of the sample. If this is adjusted for, by re-weighting the subjects according to their baseline diagnosis, crude agreement drops to 65% and Kappa to 0.53, reflecting the lower diagnostic reliability for the less advanced lesions.

Original diagnosis at baseline, age and sex were investigated as predictors of agreement using the overall diagnosis. A significant (P = 0.05) effect was found for original diagnosis but no effect was found for age and sex. The combined analysis using the results for all five biopsies produced the same result (P = 0.02 for original diagnosis). In addition, biopsy site and biopsy quality were investigated as predictors of agreement but were not found to have a significant effect.

Intra-observer Agreement

In Table 2 the overall diagnoses at first and at second gastroscopy are presented for the 45 subjects in group 2,
assessed by the same pathologist. Crude agreement is 55.6% and Kappa is 0.32. Looking at individual biopsy sites, Kappa values for biopsies A, B, C, D and F are 0.37, 0.10, 0.39, 0.30 and 0.24, respectively. Poor agreement was found for SG and AG (repeat frequency 0 and 0.28 respectively), while agreement was fair for AG, IM and Dys (repeat frequency 0.67, 0.63 and 0.70 respectively).

Age, sex and the diagnosis from the first examination were investigated as predictors of agreement between the two diagnoses, but no significant effects were found. The combined analysis using the results of all five biopsies found significant independent effects of first diagnosis \( (P = 0.001) \) and sex \( (P = 0.04) \) with histological assessment for females having higher reliability than that for males. No effects were found for age or biopsy site.

Comparison with Baseline Diagnosis

Of the 49 subjects in group 1, 17 were diagnosed at baseline by pathologist X, 16 by Y and 16 by a third pathologist (Z) who did not take part in this substudy. The results of the comparison between the routine baseline diagnosis and diagnosis at review are given in Table 3. The main determinant of agreement appears to be the identity of the pathologist who made the diagnoses at baseline. Diagnoses by pathologist Y at baseline were less frequently confirmed at review both by pathologist X and pathologist Y.

The apparent difference between pathologists at baseline could be due to confounding by histology, since agreement is higher for more advanced lesions. This possibility was investigated using a logistic regression model with agreement between baseline diagnosis and diagnosis at review as an outcome variable. A separate model was fitted for pathologist X and pathologist Y. A test for the effect of pathologist at baseline controlling for diagnosis at baseline gave \( P \)-values \( P = 0.07 \) for pathologist X and \( P = 0.01 \) for pathologist Y.

DISCUSSION

The assessment of misclassification in the diagnosis of precancerous lesions is relevant to cancer prevention trials in which early markers of neoplastic transformation are used as surrogate outcomes in place of cancer occurrence. This strategy is quite common. Out of 91 prevention trials on various cancer sites ongoing or published up to 1994, 52 used the baseline diagnosis of some precancerous lesion as an entry criterion: 28 of these used precancerous lesion changes as outcome and 15 used precancerous lesion recurrence.12

In this chemoprevention trial, precancerous lesions are diagnosed by histology using biopsies taken according to standard protocols, which is commonly considered the best criterion for defining precancerous lesions. In the absence of a gold standard measurement with which to investigate validity of the routine diagnosis, repeatability studies have been used to investigate the various components of misclassification—observer bias, random observer error and biopsy sampling error.

Agreement between the two main pathologists involved in the trial when evaluating the same samples was substantial for the overall diagnosis (Kappa = 0.69) and almost perfect for severe lesions (repeat frequency 0.96 for IM and 1.0 for Dys). In fact the original diagnosis was the main determinant of agreement. This result suggests that, when the two pathologists were involved in a review exercise, their diagnostic criteria were quite similar and misclassification was essentially limited to early lesions. Higher agreement for the more advanced lesions is consistent with the finding that agreement is higher for the overall diagnosis than for the diagnoses of individual biopsies, since the overall diagnosis is based on the most advanced lesion.

When intra-observer variation and biopsy sampling error were considered together, in a substudy in which two biopsies, separated by less than 2 months, were rated by the same pathologist, the agreement was still acceptable for the most severe lesions (repeat frequency = 0.63 for IM and 0.70 for Dys) but very low for SG and AC. It is difficult to say which of the two components of misclassification is more important. However, agreement was poor for the less advanced lesions, for which biopsy sampling error is expected to be less important due to their diffuse nature. Further, it cannot be excluded that the time elapsing between the two biopsies may have been sufficient to allow some changes in the gastric mucosa.
When routine diagnosis from three pathologists was compared with the diagnosis at review by two of them, overall agreement was quite high, but not very high. This result is expected, as a different diagnostic attitude during routine and during sample review has been described in previous studies. For one of the pathologists, however, diagnostic repeatability was low, suggesting substantial changes in diagnostic criteria from routine diagnosis to diagnosis at review.

Biopsy quality, as stated by the pathologist on the diagnostic form, proved not to be a significant determinant of agreement in this study. It should be taken into account however, that all biopsies were considered of sufficient quality to produce a diagnosis, otherwise they would not have entered in the analysis.

Overall, these substudies show that error in histological diagnosis of precancerous lesions of the stomach may be quite high. This finding has important implications for the analysis of the chemoprevention trial from which the subjects were drawn. The simplest analysis of the trial is to compare the distribution of precancerous lesions in the treatment and control groups at the end of follow-up, assuming that randomization has made the treatment and control groups comparable at baseline. However, it is also interesting to measure changes in the gastric mucosa by estimating the progression and regression rates in the two groups. In the first analysis, real differences between the two groups will be obscured by misclassification of the final diagnosis. In the second analysis, this will be combined with misclassification of the baseline diagnosis. As a result, the second analysis is expected to be less powerful, despite its greater biological interest.

The power lost due to misclassification cannot be recovered by small repeat measurement substudies, such as the ones presented in this paper. However, these results are helpful in suggesting some strategies for the analysis, such as grouping of less advanced lesions, for which misclassification is substantial. Further, they can be used as an aid in interpreting results.

ACKNOWLEDGEMENT
The authors would like to thank Colette Lavé for her work on data management.

REFERENCES

(Revised version received November 1996)