

May 2008



www.pathoindia.com

NEWSPath

Dr. S.G. Deodhare

Dr. S.G. Deodhare
Former Professor of Pathology, Grant Medical College and Dean, J.J. Group of Hospitals, Mumbai

ANATOMIC PATHOLOGY

Diagnosing Endometrial Hyperplasia: Why is it so Difficult to Agree?

Allison, Kimberly H; Reed, Susan D.; Voigt, Lynda F. et al

Current World Health Organization classification of endometrial hyperplasia is problematic because of poor diagnostic reproducibility. We sought to determine factors that cause diagnostic disagreement in a review of 2601 endometrial specimens. Blinded random specimens of normal endometrium, hyperplasias, and carcinoma were reviewed by 2 pathologists, with review by a third pathologist in cases with disagreement. All cases of endometrial hyperplasia or carcinoma were scored for degree of glandular crowding, architectural complexity, and cytologic atypia. Sample adequacy, hyperplasia volume, presence of metaplasia, or endometrial polyp were also scored. The overall [kappa] for agreement was 0.71, with a lower [kappa] of 0.36 when cases called no hyperplasia were excluded. The percent specific agreement was 90.3% for no hyperplasia, 31.1% for simple hyperplasia, 51.1% for complex hyperplasia, 49.8% for atypical hyperplasia, and 57.5% for adenocarcinoma. Cases categorized as low volume hyperplasia had more diagnostic disagreement than high volume, (62% vs. 39%, $P=0.003$). Similarly, cases called scant had more diagnostic disagreement than not scant (65% vs. 57%, $P=0.013$). The histologic feature associated with the most diagnostic disagreement was cytologic atypia ($P<0.0001$). Architectural crowding, architectural complexity, or the presence of a polyp were all associated with diagnostic disagreement ($P<0.0001$). High diagnostic disagreement in endometrial hyperplasia is related to both sample adequacy and interpretation of histologic features present. Although obtaining additional tissue may increase diagnostic reproducibility, differences in interpretation of key histologic features like cytologic atypia remain major factors contributing to diagnostic disagreement.

Various classification schemes and terminology have been applied to the endometrium, all of which have relatively poor diagnostic reproducibility. The currently accepted World Health Organization (WHO) terminology separates endometrial proliferations into simple or complex hyperplasia on the basis of architectural features and typical or atypical on the basis of cytologic features as originally defined by Kurman et al in 1985. This terminology was adopted by the WHO because of a reported increased risk of progression of lesions classified as complex hyperplasia with atypia to carcinoma in contrast to lesions diagnosed as hyperplasia without atypia (23% of cases of atypical hyperplasias progressed whereas only 2% of those without atypia progressed, within a mean follow-up of 4 y). Other studies have also reported the highest risks of progression to carcinoma in the atypical hyperplasia group, and the highest risk of persistence despite hormonal therapy. A recent prospective study also found that 43% of women with a diagnosis of atypical hyperplasia had concurrent carcinoma in hysterectomies performed within 12 weeks of diagnosis with no intervening therapy. These results suggest that a diagnosis of hyperplasia with atypia is a reliable predictor of women at high risk for subsequent or concurrent endometrial carcinoma. Currently, in the United States, a diagnosis of atypical hyperplasia usually leads to a recommendation of hysterectomy,

rather than a trial of hormonal therapy. However, how reliably and how reproducibly can pathologists make this diagnosis?

Previous studies evaluating the diagnostic reproducibility of the 1994 WHO system (the categories of which remained unchanged in the current 2003 WHO system), reported [kappa] values for overall interobserver agreement in diagnosing endometrial hyperplasia ranging from 0.2 to 0.7. The diagnosis of atypical hyperplasia was found to be the least reproducible category, with [kappa] values ranging from 0.28 to 0.65 or percent agreement as low as 25%. Various reasons for the poor reproducibility of these categories have been proposed including variably applied criteria for the diagnosis of atypia, limited samples, and complicating features such as metaplasia or polyps. But importantly, specific factors involved in diagnostic disagreement have not been systematically evaluated. In addition, some of the previous studies have suffered from inadequate documentation of methodology and low number of cases reviewed.

As part of an ongoing cohort study of 1799 women diagnosed by community pathologists with possible complex endometrial hyperplasia with or without atypia (identified through automated pathology records), we reviewed the index biopsy and subsequent endometrial samples (including normal endometrium, simple, complex, atypical hyperplasia, and carcinoma) for 2601 specimens. Diagnostic agreement was evaluated between review panel pathologists. Each case was scored for a variety of factors related to specimen quality, quantity, and diagnostic criteria to elucidate features associated with diagnostic disagreement.

MATERIALS AND METHODS

As part of the Endometrial Hyperplasia Outcomes (ECO) cohort study, 2601 endometrial specimens underwent pathology review. Women aged 18 to 88 years, with a possible diagnosis of complex endometrial hyperplasia with or without atypia, were identified through records from January 1, 1985 to April 1, 2005 from automated pathology databases at Group Health in Washington State. To maximize the number of hyperplasia specimens and mimic how patients were most likely to be clinically classified, diagnoses with indefinite wording such as cannot rule out or at least were included in the pathology review and categorized as the higher grade diagnosis, with the exception of specimens with a diagnosis of atypia cannot rule out low-grade carcinoma, which were included as possible atypical hyperplasias. Women initially diagnosed with endometrial carcinoma were excluded. Both index diagnostic specimens (1799 specimens) and all follow-up specimens (702 specimens) were included in the randomized, blinded pathology review. A total of 2418 specimens were biopsies or curettages and 183 were hysterectomies.

Each case was independently reviewed by 2 pathologists, with additional independent review by a third pathologist in cases with diagnostic disagreement. Cases were assigned a final diagnosis on the basis of the majority diagnosis. If there was no majority diagnosis, the senior pathologist (R.L.G.) reviewed the 3 diagnoses and selected the middle category diagnosis. Final diagnosis was classified as simple hyperplasia, complex hyperplasia, atypical hyperplasia, carcinoma, or no hyperplasia (specified as proliferative,

normal secretory, atrophic/inactive, or shedding/menstrual). Carcinomas were graded according to the International Federation of Gynecology and Obstetrics (FIGO) system. All cases diagnosed using WHO criteria as endometrial hyperplasia or endometrial carcinoma were scored for degree of glandular crowding, architectural complexity, and cytologic atypia. In addition, sample adequacy, volume of hyperplastic tissue, and presence of metaplasia or endometrial polyp were noted.

All 3 pathologists are academic pathologists. The 2 primary reviewers have a subspecialty focus in gynecologic pathology (K.H.A. and R.L.G.). R.L.G. and C.D.J. have each had over 10 years experience in practice and K.H.A. was a fellow in gynecologic and breast pathology at the study onset and became faculty during the study. Pathologists reviewed an initial pilot series of cases together to establish definitions for the features scored) and pathologists were instructed to use the current WHO criteria in establishing a diagnosis; thereafter cases were reviewed independently. Initial pilot studies on 38 specimens for reviewer R.L.G. resulted in no intraobserver variability. Pathologists were blind to community-based diagnosis and all clinical information with the exception of patient age. Index and follow-up biopsies were randomly mixed for review.

Information regarding demographics (age and race), reproductive, medical and family history, and physical characteristics, including height and weight at the time of the index biopsy, were collected from the Group Health medical record; a single document containing all records from outpatient visits, test reports, and records of hospitalizations and consultations.

The [kappa] statistic was used to measure interreviewer agreement. Kappa values were computed using STATA 9.2 (StataCorp, College Station, TX). Percent agreement was computed as the number of specimens where the 2 reviewers agreed exactly divided by the total number of eligible specimens. Overall weighted [kappa] and weighted agreement were computed for by ordering the diagnoses as follows: no hyperplasia, simple hyperplasia, complex hyperplasia, and hyperplasia with atypia and carcinoma. The following standard weights in STATA 9.2 were used: 1.0 for perfect agreement, 0.75 for adjacent categories, 0.5 for diagnosing 2 categories away from each other, 0.25 for those 3 categories away, and 0 for all others. Kappa values for individual diagnoses were computed by collapsing the data into 2-by-2 tables (agreed on a diagnosis, either reviewer disagreed on this diagnosis, both agreed the specimen was some other diagnosis) as described by Fleiss. Percent-specific agreement was computed for individual diagnoses instead of percent agreement overall to avoid giving inappropriate weight to the large category in each table where both reviewers agreed the specimen was some other diagnosis. P values for the difference between proportions were computed using the PEPI program Differ. In the data analysis for Tables 1 to 3, cases where both reviewers diagnosed no hyperplasia or either reviewer diagnosed cannot rule out hyperplasia, inadequate sample, or nondiagnostic were excluded because they were not scored for volume of hyperplasia or the presence of specific diagnostic criteria.

	Cases With Diagnostic Disagreement	
	N	%
Specimen adequacy		
"Scant" (< 0.5 cm ³), N = 353	228	64.6
Not scant, N = 791	447	56.5
<i>P value of difference</i>	0.013	
Volume of hyperplasia		
Low volume (< 0.5 cm ³), N = 761	471	61.9
High volume (> 0.5 cm ³), N = 227	88	38.8
<i>P value of difference</i>	< 0.0001	
Type of specimen		
Biopsy or curettage, N = 1060	629	59.3
Hysterectomy, N = 101	46	55.5
<i>P value of difference</i>	0.51	

TABLE 1. Correlation of Quantity of Diagnostic Material With Diagnostic Disagreement

	Disagreed on Cytologic Atypia		Disagreed on Degree of Glandular Crowding		Disagreed on Architectural Complexity	
	N	%	N	%	N	%
Cases with diagnostic agreement, N = 417	67	16.1	88	21.1	92	22.1
Cases with diagnostic disagreement, N = 364	172	47.3	123	33.5	115	31.3
<i>P</i>	< 0.0001		< 0.0001		0.005	

TABLE 2. Histologic Features Involved in Diagnostic Disagreement

Study	N Total	Case Selection	Reviewers	Overall κ	NH	SH	CH	AH	CA
Skov et al ²⁶	128	Review of cases submitted as hyperplasia	2 Institutions in Denmark, 6 pathologists	0.20-0.25	NA	0.21-0.25	0.07-0.15	0.5-0.65	NA
Kendall et al ¹¹	100	Review of a selected variety of diagnoses	Single US institution, 5 pathologists	0.7	0.9	0.6		0.5	0.8
Bergeron et al ²	56	Review of a selected variety of diagnoses	Multiple European institutions, 5 pathologists	0.47	71%	44%	18%	25%	49%
Zaino et al ²⁵	302	Review of cases submitted as atypical	Multiple US hospitals, 3 pathologists	0.4	0.48	0.38		0.28	0.51
Allison, 2007	2,601 (1,161 "not normal")	Women with a possible diagnosis of hyperplasia	Single institution, 3 pathologists	0.71 91% 0.36 (for hyperplasia only)	0.76 90%	0.16 31%	0.21 51%	0.35 50%	0.55 58%

Kappa: 0.0-0.2 = slight agreement, 0.21-0.40 = fair, 0.41-0.60 = moderate, 0.61-0.80 = substantial, 0.8-1.00 = almost perfect agreement.¹⁷
 AH indicates atypical hyperplasia; CA, carcinoma; CH, complex hyperplasia; NH, no hyperplasia; SH, simple hyperplasia; %, % agreement.

TABLE 3. Comparison of Studies on 1994 WHO Endometrial Hyperplasia Classification Scheme Reproducibility Kappa: 0.0-0.2=slight agreement, 0.21-0.40=fair, 0.41-0.60=moderate, 0.61-0.80=substantial, 0.8-1.00=almost perfect

agreement. AH indicates atypical hyperplasia; CA, carcinoma; CH, complex hyperplasia; NH, no hyperplasia; SH, simple hyperplasia; %, % agreement

RESULTS

A total of 1799 women were included in the study cohort population. Clinical characteristics of the cohort are summarized in Table 4. The most common age range was 45 to 54 years. Most of the women were white and it was most common to have a body mass index of 30 kg/m² or greater.

The 2 primary pathologists reviewed a total of 2601 specimens with 577 (22.2%) resulting in diagnostic disagreement, which went to review by the third pathologist. Final panel diagnosis resulted in 1829 no hyperplasia or simple hyperplasia diagnoses, 396 complex hyperplasia, 288 atypical hyperplasia, 54 adenocarcinoma, and 34 other (inadequate, nondiagnostic or cannot rule out hyperplasia) (Table 5).

Overall diagnostic agreement between the 2 main panel pathologists was 91.1% (weighted percent agreement), with a weighted [kappa] of 0.71 (Table 6). Disagreements resulting in a third review were most frequent when one pathologist diagnosed complex hyperplasia and the other diagnosed complex hyperplasia with atypia (29.2% of cases with disagreement). Most disagreements were within one diagnostic category of each other, with a small percent (9.3%) of disagreements significant up or downgrades from no hyperplasia or simple hyperplasia to atypical hyperplasia or carcinoma or vice versa.

	N	%
Age (y)		
< 45	288	16.8
45-54	752	43.9
55-64	379	22.1
≥ 65	293	17.1
Race		
White	1455	85.0
Black	44	2.6
Asian	75	4.4
Other	29	1.7
Unknown	109	6.4
Body mass index (kg/m²) at reference date		
Not overweight (< 24.9)	531	32.9
Overweight (25-29.9)	406	25.2
Obese (30-39.9)	426	26.4
Morbidly obese (≥ 40)	251	15.6
Missing	98	

TABLE 4. Clinical Characteristics of Cohort

Final Diagnosis	Frequency	Percent
No hyperplasia or simple hyperplasia	1829	70.3
Complex hyperplasia	396	15.2
Atypical hyperplasia	288	11.1
Adenocarcinoma	54	2.08
Other*	34	1.31
Total	2601	

*Inadequate specimen, nondiagnostic or cannot rule out hyperplasia.

TABLE 5. Frequencies of WHO Final Panel Diagnoses*Inadequate specimen, nondiagnostic or cannot rule out hyperplasia.

	Percent-specific Agreement	κ
No hyperplasia, N agreed = 1370	90.3	0.76
Simple hyperplasia, N agreed = 67	31.1	0.16
Complex hyperplasia, N agreed = 226	51.1	0.21
Atypical hyperplasia, N agreed = 139	49.8	0.35
Adenocarcinoma, N agreed = 44	57.5	0.55
Overall except “no hyperplasia,” N total = 1161	80.6*	0.36*
Overall, N total = 2531	91.1*	0.71*

*Weighted percent agreement or weighted κ.

TABLE 6. Diagnostic Agreement by Diagnosis*Weighted percent agreement or weighted [kappa].

Diagnostic trends between the 2 primary pathologists are shown in Figure 1A. Pathologist A was more likely to diagnose atypical hyperplasia or carcinoma whereas pathologist B was more likely to diagnose complex, simple, or no hyperplasia. But overall diagnostic trends were similar. Diagnostic trends for cases sent to pathologist C are shown in Figure 1B. Reviewer C agreed with 43.9% of pathologist B's and 27.7% of pathologist A's diagnoses. Reviewer C also had a similar frequency of diagnosing carcinoma as reviewer B. However, reviewer C diagnosed atypical hyperplasia more often than reviewer B.

Diagnostic agreement by the panel for each WHO category is shown in Table 6. Agreement was highest for no hyperplasia and lowest for simple hyperplasia. Agreement for complex and atypical hyperplasia was fair ($[\kappa]=0.21$ and 0.35), and agreement for adenocarcinoma was moderate ($[\kappa]=0.55$). There were no statistically significant differences in agreement for index verses follow-up specimens.

Association of the quantity of diagnostic material present in each case with diagnostic disagreement is presented in Table 1. Cases with specimen adequacy scored by either pathologist as scant had more diagnostic disagreement than those scored moderate or abundant ($P=0.013$). In addition, a low volume of hyperplastic tissue present was significantly associated with diagnostic disagreement with ($P<0.0001$). Cases with a high volume of hyperplasia had the least amount of diagnostic disagreement (38.8%). Nonhysterectomy and hysterectomy specimens had similar diagnostic disagreement ($P=0.51$).

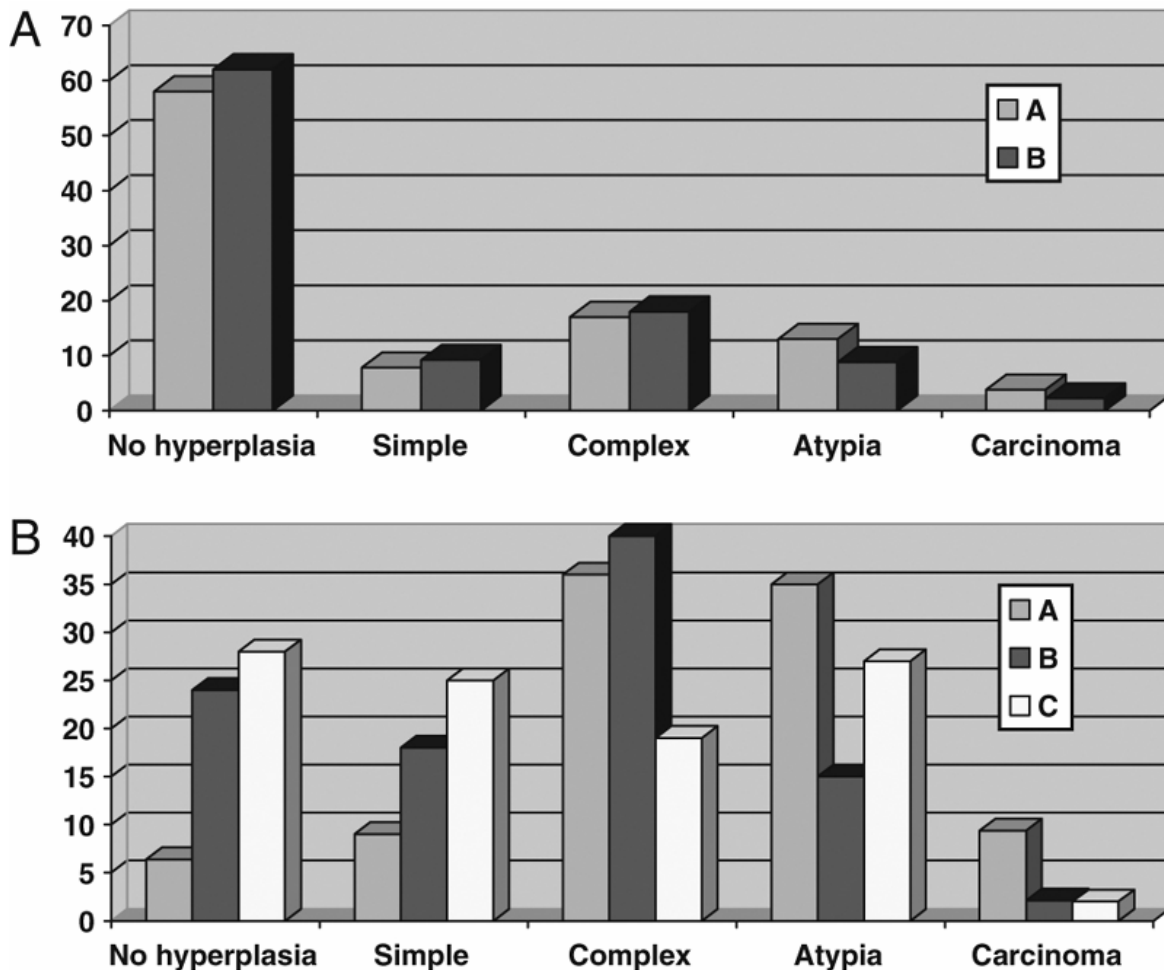


FIGURE 1. A, Diagnostic trends of the 2 primary panel pathologists for all specimens. Pathologist A was more likely to diagnose atypical hyperplasia or carcinoma whereas pathologist B was more likely to diagnose complex, simple, or no hyperplasia. However, overall diagnostic trends were similar. B, Diagnostic trends

for cases sent to third reviewer. The third reviewer, C, agreed with 43.9% of pathologist B's and 27.7% of pathologist A's diagnoses. The third reviewer diagnosed atypical hyperplasia more often than reviewer B (but less than A), no hyperplasia or simple hyperplasia more frequently and complex hyperplasia less frequently than both reviewers A and B. Reviewer C also had a similar frequency of diagnosing carcinoma to reviewer A.

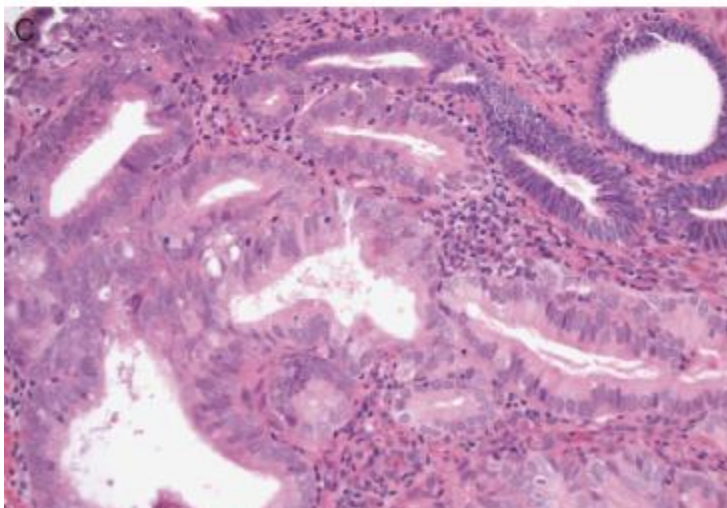
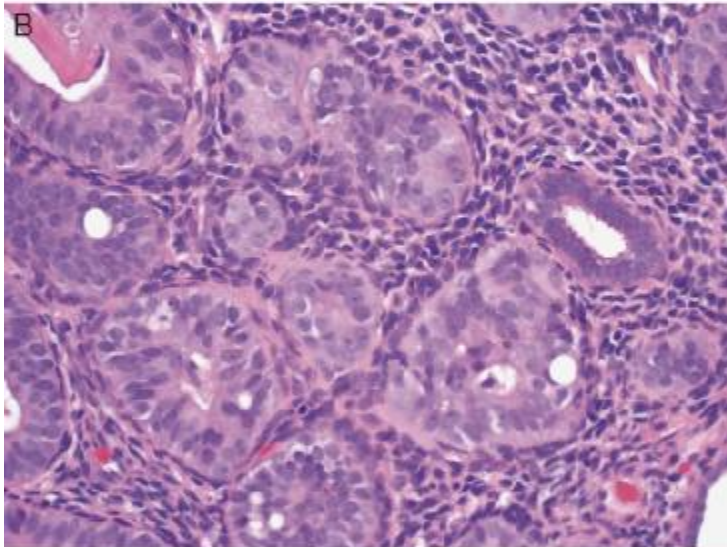
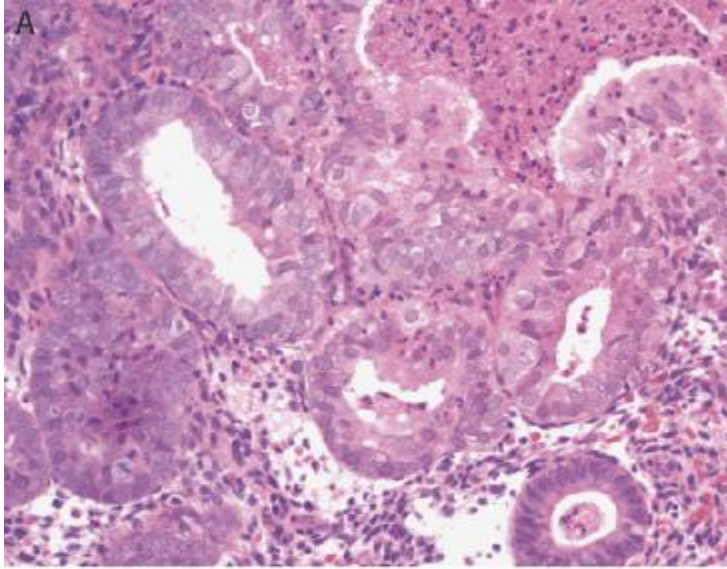


FIGURE 2. Examples of the histology of cases with agreement and disagreement on the presence of cytologic atypia. A, Case with agreement on both the presence of cytologic atypia and the diagnosis of atypical hyperplasia. B, Case that was quantified as scant and low volume neoplasia with disagreement about the presence of cytologic atypia. The case had a final classification as atypical hyperplasia. C, Case with disagreement on the presence of cytologic atypia, which had a final classification as complex hyperplasia without atypia. The most atypical appearing focus is pictured, which was very focal. Original magnification —200 (A).

Cases with diagnostic disagreement were more likely to disagree on all 3 histologic features evaluated than cases with diagnostic agreement (Table 2). Almost half (47.3%) of the cases with diagnostic disagreement had disagreement on the degree of cytologic atypia, compared with 16.1% of the cases with diagnostic agreement ($P<0.0001$). Cases with disagreement about the degree of glandular crowding included 33.5% of cases with diagnostic disagreement, compared with 21.1% of the cases with diagnostic agreement ($P<0.0001$). Cases with disagreement about the degree of architectural complexity included 31.3% of cases with diagnostic disagreement, compared with 22.1% of the cases with diagnostic agreement, ($P=0.005$). See Figure 2 for the histology of examples of cases with diagnostic disagreement.

Complicating factors such as the presence of a polyp or abundant metaplasia were also investigated for their effect on agreement. There was greater diagnostic disagreement when features of a polyp were noted by one of the pathologists. Of 230 cases where a polyp was noted to be present, 164 (71.3%) also had diagnostic disagreement compared with 521 of 931 cases (56%) where no polyp was noted ($P<0.0001$). However, there was only a suggestion of increased diagnostic disagreement when metaplastic changes were noted by one of the pathologists (66.2% of cases noted to have metaplasia had diagnostic disagreement compared with 58% of cases not noted to have metaplasia, $P=0.083$).

DISCUSSION

With 2601 specimens reviewed, to our knowledge, this study is the largest blinded review of the reproducibility of the 1994 WHO endometrial hyperplasia classification scheme to date. Our overall agreement ($[\kappa]=0.71$), was similar to other studies which have $[\kappa]$ values ranging from 0.2 to 0.7, Table 3. Our agreement was lowest for the diagnosis of simple hyperplasia ($[\kappa]=0.16$) and highest for not hyperplastic endometrium ($[\kappa]=0.76$). Because of the high percentage of specimens with a final diagnosis of no hyperplasia, our overall agreement was inflated by our high agreement for nonhyperplastic specimens. In fact when no hyperplasia cases were excluded our overall agreement was significantly lower ($[\kappa]=0.36$, specific agreement 80.6%). However, one might argue that having high numbers of nonhyperplastic specimens admixed with a variety of hyperplastic specimens is more analogous to actual clinical practice settings.

Although previous studies are not directly comparable due to differences in initial case selection and review panels, our agreement for specific diagnostic categories was similar both Zaino's 2006 prospective Gynecologic Oncology Group (GOG) study and Bergeron's 1999 European multi-institution study for the categories of nonatypical hyperplasia, atypical hyperplasia, and adenocarcinoma (Table 3). With $[\kappa]$ values for agreement on atypical hyperplasia, a key diagnostic clinical decision point, as low 0.2 to 0.3 for expert gynecologic pathologists, there is clearly an issue with the reproducibility of the current WHO diagnostic scheme.

With specialists having so much trouble agreeing, it is not surprising that there is frequent disagreement between specialists and community pathologists. In Zaino's GOG study the majority review panel diagnosis supported the referring institution diagnosis in only 38% of cases submitted as atypical hyperplasia. In our study, the final review panel diagnosis and the initial outside diagnostic categorization were not directly comparable, because the method of initial diagnosis categorization was intended to maximize the number of possible cases of complex and atypical hyperplasia selected for review (see Methods). Given our bias toward categorization of the initial diagnosis to a higher grade diagnostic category, it was not surprising that we had trends toward down-grading the initial diagnostic category by the review panel final diagnosis (data not shown). However, for the above reasons, agreement with the original diagnosis was not considered of value in this study.

What are the factors that cause diagnostic disagreement? Our study is the first of its kind to systematically investigate the contribution of sample adequacy, interpretation of key histologic features, and the presence of complicating features (polyps and metaplasias) to diagnostic disagreement.

Clearly, problems with adequate sampling are an issue beyond effecting diagnostic agreement, with other studies showing the rates of finding concurrent carcinoma in hysterectomy specimens with a review panel diagnosis of normal or nonatypical hyperplasia as high as 19%. But, with a given amount of diagnostic tissue, how do the amount of total tissue present for evaluation and the total amount of hyperplastic tissue present effect diagnostic agreement? In our study, specimens categorized by either pathologist as scant ($<0.5 \text{ cm}^3$) or low volume of hyperplastic tissue were significantly more likely to have disagreement about the diagnosis ($P \leq 0.013$ and $P \leq 0.0001$, respectively). This implies that specimens that have either a minimal amount of diagnostic tissue total (sample borders on inadequate), or samples that have only a very focal amount of hyperplastic tissue in otherwise normal endometrium, should be reviewed with caution. These samples may warrant a comment about the small amount of diagnostic tissue present and the uncertainty in the diagnosis and request additional sampling. Zaino et al suggest similar findings related to sample adequacy, with greater diagnostic reproducibility for dilation and curettage specimens than office biopsy or curettage methods. Interestingly, we did not find a statistically significant difference in diagnostic agreement between hysterectomy and nonhysterectomy specimens, possibly because even hysterectomies can have very low volumes of hyperplastic tissue, which can decrease diagnostic agreement.

Although sampling is an issue that can be controlled by recommending additional tissue, the lack of objectivity in applying multiple diagnostic criteria to establish a diagnosis is more challenging. The histologic features referred to in the WHO as useful in establishing a diagnosis include architectural changes, shift in the gland to stoma ratio, and cytologic atypia. However, strict definitions of these features and the criteria used to establish a specific WHO diagnosis are not spelled out in great detail. Architectural changes said to be characteristic of complex hyperplasia include irregular epithelial budding and increased gland complexity, which is not further defined. A shift in gland to

stroma ratio in favor of the glands is noted by the WHO to be a feature of complex hyperplasia as well but a strict threshold is not set. The endometrial intraepithelial neoplasia (EIN) scheme used by Mutter and colleagues is more specific, using a volume percent stroma of less than 55% (area of glands > stroma) as one of the diagnostic criteria for a diagnosis of EIN. But it is the subjective interpretation of the presence of cytologic atypia in the WHO scheme that seems to be most problematic. In fact, the WHO specifically states definitions of cytologic atypia are difficult to apply in the endometrium because nuclear cytologic changes occur frequently in hormonal imbalance, benign regeneration, and metaplasia. The WHO describes nuclear rounding, loss of polarity, prominent nucleoli, irregular nuclear membranes, and cleared or dense chromatin as features of cytologic atypia but acknowledges that atypia may be best observed by comparison with the adjacent normal glands. The EIN scheme avoids using a descriptive definition of cytologic atypia and instead uses distinct cytology in the architecturally crowded focus that is different from background. Given the fairly loosely defined WHO diagnostic criteria, we were interested in determining if disagreement about the presence of key histologic features was a major factor in whether there was agreement about a specific WHO diagnosis.

Other studies have evaluated which histologic features could most aid recognition of cytologic atypia or architectural complexity. Kendall et al found gland crowding significantly associated with a diagnosis of complex hyperplasia whereas nucleoli was the only feature significantly associated with a pathologist calling a case atypical. Bergeron et al also found the presence of gland crowding most significantly associated with a diagnosis of hyperplasia whereas nuclear pleomorphism was most significantly associated with classification as atypical. However, others have not investigated how concordance on the presence of certain histologic features specifically effect diagnostic agreement. Because our study did not include outcomes, we did not intend to define which features were more predictive of risk of carcinoma, but merely to investigate if we could agree on the presence of defined features and if disagreement of their presence effected agreement on final diagnosis.

In our study, cases with diagnostic disagreement were also more likely to disagree on specific key histologic features such as architectural complexity, glandular crowding, and cytologic atypia, than cases with diagnostic agreement, indicating variable application of defined histologic features to formulate a diagnosis. Cytologic atypia was the feature most often disagreed on and had the largest difference between cases with disagreement versus agreement (47.3% of cases with diagnostic disagreement also disagreed on the presence of cytologic atypia, vs. 16.1% of cases with diagnostic agreement). Although our study only reflects the agreement between 2 pathologists, the poor reproducibility of atypical hyperplasia in previous studies supports these findings. Given that the presence of atypia is currently considered the best predictor of outcome in the WHO scheme, this finding calls into question the reliability of using atypia as it is currently defined (and variably interpreted) as a breakpoint for diagnostic categories.

The final factors we investigated as possible causes of diagnostic disagreement were complicating histologic features the presence of features of a polyp or metaplasia.

Because polyps tend to be less hormone responsive they can have more irregularly distributed glands with areas of crowding and have various metaplastic cytologic changes that can make differentiation of normal polyp from a polyp with areas of hyperplasia challenging. We did find a significant association with diagnostic disagreement in specimens where either pathologist had noted there were features of a polyp present (N=230, P<0.0001). Better criteria are needed to distinguish changes in polyps that should be considered higher risk, neoplastic lesions. In addition, when crowding or cytologic changes are limited to a polyp, a comment as to the unclear significance of the changes may be warranted.

The presence of metaplastic changes, or epithelial cytoplasmic change, in the endometrium varies from squamous, to tubal, to repair-associated eosinophilic syncytial change. The presence of extensive metaplasia can complicate the diagnosis of hyperplasia by making glands look more crowded (especially in extensive squamous metaplasia) or cytologically atypical. To complicate matters further, metaplastic changes are often associated with hyperplasias. We did note greater diagnostic disagreement when either pathologist noted the presence of metaplasia, however, this did not reach statistical significance (P=0.083). This may have occurred because noting the presence of metaplasia was an optional part of the scoring form. In fact, of the 10 cases called cannot rule out hyperplasia, the most common reason noted was extensive metaplastic changes. Metaplastic changes are histologic features to be aware of as a possible pitfall in diagnosing endometrial hyperplasia but, while metaplasia was commonly associated with a diagnosis of cannot rule out hyperplasia, it was not a major cause of diagnostic disagreement in this study.

Additional studies to establish more reproducible criteria for endometrial hyperplasia that are also predictive of progression to carcinoma are needed. Various alternate diagnostic schemes have been proposed. Bergeron et al proposed combining simple and complex hyperplasia into a single Hyperplasia group and combining atypical hyperplasia with a subset of well-differentiated carcinomas into an endometrial neoplasia group. This scheme has the advantage of combining lesions that are treated in a similar way but its diagnostic utility has not been investigated. Mutter and colleagues have more thoroughly investigated another scheme that was developed from molecular, histomorphometric, and outcome data which separates precancerous neoplastic Endometrial intraepithelial neoplasia from benign endometrial hyperplasia due to the presumed influence of unopposed estrogens. However, while this system has the advantage of strong correlation of EIN with clonal populations, it is still unclear if this broad category can be further refined into high-risk neoplasms that are likely to persist or progress to invasive carcinomas despite treatment with progestins versus lower risk neoplastic populations that may be spontaneously shed or regress with progestin therapy.

In conclusion, our study, the largest of its kind to date, confirms previous findings related to the poor reproducibility of the current WHO endometrial hyperplasia classification system. In addition, our findings suggest that diagnostic disagreement is due both to inability to agree on the presence of various key histologic features and the amount of diagnostic tissue present. We suggest that in the clinical setting specimens with limited

amounts of diagnostic tissue (either low volumes of diagnostic hyperplastic tissue or overall scant specimens) should be interpreted with caution and that recommending additional tissue should be considered. Setting a threshold for amount of diagnostic tissue present necessary for a definitive diagnosis could improve diagnostic agreement and decrease the rates of immediate hysterectomies for this usually low-grade neoplastic process. Given the poor reproducibility of the diagnosis of endometrial hyperplasia, studies examining outcomes in this field may also want to consider limiting their cases to those that have diagnostic agreement among reviewing pathologists or at least have a minimum threshold of diagnostic tissue present. Using stricter criteria for outcome studies will help give a clearer picture of the natural history of endometrial hyperplasia and perhaps shed more light on which lesions are truly higher risk.

The American Journal of Surgical Pathology, Volume 32(5), May 2008, pp 691-698

Cirrhosis-associated Hepatocellular Nodules: Correlation of Histopathologic and MR Imaging Features

Robert F. Hanna, Diego A. Aguirre, Norbert Kased et al

Cirrhotic livers are characterized by advanced fibrosis and the formation of hepatocellular nodules, which are classified histologically as either (a) regenerative lesions (eg, regenerative nodules, lobular or segmental hyperplasia, focal nodular hyperplasia) or (b) dysplastic or neoplastic lesions (eg, dysplastic foci and nodules, hepatocellular carcinomas). The differentiation of these lesions is important because regenerative nodules are benign, whereas dysplastic and neoplastic nodules are premalignant and malignant, respectively. However, their accurate characterization may be difficult even at histopathologic analysis. Differential diagnosis may be facilitated by comparing the clinical and pathologic findings with radiologic imaging features; in particular, nodule size, vascularity, hepatocellular function, and Kupffer cell density assessed at magnetic resonance (MR) imaging are suggestive of the correct diagnosis. MR imaging is more useful than computed tomography for such assessments because it provides better soft-tissue contrast and a more nuanced depiction of different tissue properties. Moreover, a wider variety of contrast agents is available for use in MR imaging. Familiarity with the MR imaging characteristics of cirrhosis-associated hepatocellular nodules is therefore important for optimal diagnosis and management of cirrhotic disease.

RadioGraphics 2008;28:747-769

CLINICAL PATHOLOGY

Best Practice in Primary Care Pathology: Review 11

Abstract

This eleventh best practice review examines common primary care question in laboratory medicine: thyroid testing. The review is presented in the same question answer format as in the previous reviews. These questions and answers deal with common situations in men and non-pregnant women. The recommendations represent a guidance found using a standardised literature search of national and international guidance notes, consensus statements, health policy documents and evidence-based medicine reviews, supplemented by Medline Embase searches to identify relevant primary research documents. In the case of the thyroid series, the recommendations are drawn from the 2006 guidelines published by the Association for Clinical Biochemistry, the British Thyroid Association and the British Thyroid Foundation. They are not standards but form a guide to be set in the clinical context. Most are consensus rather than evidence based. They will be updated periodically to take account of new information.

THYROID TESTING

UK guidelines for the use of thyroid function tests were published in June 2006 jointly by the Association for Clinical Biochemistry, the British Thyroid Association and the British Thyroid Foundation. These guidelines have provided a comprehensive literature review and evidence-based guidelines for the rational use of thyroid function tests for the diagnosis and management of thyroid disorders. They were written to encourage a greater understanding of thyroid function testing amongst all stakeholders including laboratory personnel, clinicians in primary and secondary care and patients and their carers. In drawing up these guidelines it was clear that there was a lack of high-quality evidence in clinical thyroid disease in the form of randomised controlled trials and meta-analyses. Consequently there was much reliance on second-level evidence such as cohort and case control studies, with good practice points often used to plug the gaps where no real evidence existed. The guidelines, which contain over 200 recommendations, cover all aspects of thyroid disease including indications for thyroid function testing, hypothyroidism, hyperthyroidism, pregnancy, thyroid cancer and laboratory aspects of thyroid function testing. In order to illustrate the use of these guidelines, a common scenario presenting in primary care is described.

Which thyroid function test combination should a laboratory provide?

- Serum thyroid stimulating hormone (TSH) is considered a suitable screening test for primary hypothyroidism in most patients with additional tests (free thyroxine (FT4)) if outside the reference interval.
- If TSH is to be used alone, users and laboratories must identify patients thought to have possible pituitary hypothyroidism so that TSH and FT4 are measured in these cases.
- Serum TSH alone is sufficient in follow-up testing of patients at risk of developing hypothyroidism who are not being treated for thyroid disorders.

A strategy of first-line TSH may be cost effective for a wide range of clinical purposes including screening and case finding, but it may be inappropriate in patients being tested for the first time, and in some specific clinical settings. The guideline stresses that pituitary hypothyroidism can produce TSH concentrations within the reference interval but with low FT4. If laboratories are unable to identify those specimens that specifically require the measurement of serum TSH and FT4 then it would be prudent to measure serum TSH and FT4 on all specimens rather than embark on a first-line serum TSH strategy.

This would have considerable cost implications for primary care users, and we therefore recommend in the first instance that users should highlight cases in which any suspicion of hypothyroidism of pituitary origin are suspected. Measurement of serum TSH alone is appropriate after the first investigation in the sequential follow-up of individuals who have not been treated for thyroid disorders and who may be at risk of developing primary thyroid dysfunction.

In which patients should thyroid testing be performed systematically?

A 33-year-old woman presents to her family doctor complaining of gaining weight (5 kg in 6 months) and feeling tired all the time. There is no relevant family history and examination is normal.

We recommend the following.

- Thyroid testing (TSH and FT4 if TSH outside reference interval) is recommended in all patients presenting with goitre or thyroid nodule, atrial fibrillation, osteoporosis, subfertility or dyslipidaemia (particularly if total cholesterol >8 mmol/l).
- Thyroid testing (TSH, FT4 and anti-thyroperoxidase (anti-TPO) antibodies) is recommended pre-conception, at booking and at 3 months post partum in all women with type 1 diabetes.
- Patients with type 1 diabetes should have annual TSH testing, and those with type 2 diabetes should be tested at the time of diagnosis.
- Annual TSH screening (and screening before and 6-8 weeks after subsequent pregnancies) is recommended in all women who have suffered post-partum thyroiditis.
- Annual TSH screening is recommended in all patients with Down or Turner syndrome.
- Patients taking specific drugs (eg, amiodarone, lithium) should be tested according to established guidelines.

Targeted thyroid function testing is recommended on the basis of the relatively high prevalence of thyroid dysfunction in selected groups. In particular, thyroid function testing at diagnosis is considered to be cost-effective in type 2 diabetic patients. Routine testing in patients acutely admitted to hospital is not recommended because of the high

prevalence of non-thyroidal illness, producing lowered or raised TSH concentrations in the absence of thyroid disease.

Should opportunistic population thyroid function testing be performed?

- Screening of the healthy population is not recommended.
- Opportunistic screening of adult women at menopause or presenting in primary care with non-specific symptoms may be considered.

Recommendations on population screening for thyroid disease, particularly subclinical hypothyroidism, vary. Targeted opportunistic screening in menopausal women or women with non-specific symptoms is recommended based principally on a cost-effectiveness analysis, suggesting that screening for hypothyroidism in this group compares well with other preventative practices and improves quality of life.

TSH within reference interval, low/normal FT4: does this patient have primary hypothyroidism and require treatment with thyroxine?

A GP receives the thyroid function test result FT4 10.0 pmol/l (reference interval 10-20 pmol/l) and TSH 4.4 mU/l (reference interval 0.4-4.5 mU/l) and contacts the laboratory for advice.

We recommend the following.

- The diagnosis of primary hypothyroidism requires the measurement of TSH and FT4.
- Primary hypothyroidism is excluded if the serum TSH is in the reference interval and the patient is not taking medication known to affect TSH.
- Consider secondary (pituitary or hypothalamic) causes if TSH is in the reference interval but FT4 is reduced.
- Note that reference intervals may differ in pregnant women depending on trimester of pregnancy.

It is recognised that there is a growing distrust of thyroid function tests among a minority of patients because they have test results within the reference interval but have symptoms suggestive of hypothyroidism. It is accepted that the guidelines quote decision limits for TSH with the aim of simplifying, standardising and optimising clinical decisions. It is recognised that there can be variability in bias between the various commercial assays available. Nevertheless, large population surveys from the US using rigorous criteria for selecting a healthy reference population have failed to provide any evidence for narrowing the reference interval of TSH.

Raised TSH, low/normal FT4: when should evidence of thyroid failure be treated?

The GP decides to repeat the test 6 months later. The symptoms have persisted. The following thyroid function test results are now available: FT4 11.2 pmol/l (reference

interval 10-20 pmol/l) and TSH 8.5 mU/l (reference interval 0.4-4.5 mU/l) and the GP contacts the laboratory for advice.

We recommend the following.

- If serum TSH on screening is mildly raised (5-10 mU/l), and the FT4 is within the reference interval:
 - exclude non-thyroidal illness and drug interference
 - repeat 3-6 months later, with measurement of serum FT4
 - measure serum anti-TPO antibodies.
- If the serum antibody measurement is positive:
 - measure serum TSH annually or earlier if symptoms
 - start thyroxine therapy if the serum TSH rises above 10 mU/l
 - consider possible trial of thyroxine if TSH over 5 mU/l on individual case basis.
- If the serum antibody measurement is negative:
 - repeat measurement of serum TSH approximately every 3 years.
- If the serum TSH is >10 mU/l and serum FT4 concentration is within the reference interval:
 - the person has overt hypothyroidism and should be treated with thyroxine.
- If the serum TSH concentration is >10 mU/l and serum FT4 concentration is within the reference interval:
 - treatment with thyroxine is recommended in most cases
 - further advice should be sought if FT4 well within normal range (eg, >15 pmol/l).
- If the serum TSH concentration is above the reference interval but <10 mU/l:
 - there is no evidence to support the benefit of routine early treatment with thyroxine in non-pregnant patients with a serum TSH above the reference interval but <10 mU/l. A therapeutic trial of thyroxine may be considered on an individual patient basis.

The guideline recommends that patients with a TSH greater than 10 mU/l and FT4 below the reference interval should be considered to have overt hypothyroidism. In those with TSH between 5 and 10 mU/l and FT4 within the reference interval, a further test in 3-6 months is recommended, combined with thyroid antibody (TPO antibody) measurement. Positive anti-TPO antibodies (above the positivity threshold for the laboratory method used) indicate a high likelihood of developing hypothyroidism. Treatment is recommended in patients with TSH greater than 10 mU/l but normal FT4 concentration. There is no clear statement on action in patients with TSH between 5 and 10 mU/l and low FT4 and it would appear reasonable to decide on an individual case basis whether to offer treatment or continue to monitor depending on symptoms and actual FT4 and TSH values. Patients with positive TPO antibody are recommended to have an annual thyroid function test. Those who are antibody negative are recommended to have a thyroid function test every 3 years.

What should be the target for TSH in those on thyroxine replacement?

The hypothyroid patient above is positive for TPO antibody. After 2 years the serum TSH has risen to 12 mU/l. She is commenced on thyroxine replacement and stabilised on a dose of 100 µg daily. Her symptoms have persisted. When tested 8 weeks later, the following thyroid function test result is available: FT4 19.4 pmol/l (reference interval 10-20 pmol/l) and TSH 0.5 mU/l (reference interval 0.4-4.5 mU/l).

We recommend the following.

- Annual measurement of thyroid function (minimum of TSH) is recommended in all patients receiving long-term replacement thyroxine.
- Measurement of TSH and FT4 is usually recommended to assess thyroid replacement with thyroxine. TSH alone may be all that is required and measurement of FT4 may only be required if TSH is outside of the reference interval.
- Replacement should be assessed from clinical well-being and thyroid testing.
- A minimum interval of 2 months is recommended before measuring thyroid function after changing thyroxine dose.
- Thyroid measurement may be appropriate 2 months after starting drugs which influence thyroxine requirements (box 1).
- Recommended aims in thyroid replacement are for TSH within and FT4 within (or slightly above) the population reference interval.
- Not all patients will be clinically optimally controlled within this range.

The guidance is based on observational studies indicating high rates of subclinical hypo- and hyperthyroidism in patients taking long-term thyroxine, and on the potential effect of some commonly prescribed drugs (iron salts, oestrogens, phenytoin, carbamazepine) to alter thyroxine requirements. This could suggest benefit from monitoring after starting such drugs. The guideline recommended that variations in dosage requirement due to concomitant drugs be taken into account. After changing dose, thyroid function should not normally be measured within 2 months, the period required to reach thyroid steady state. The recommended target range is based on TSH, for which most evidence is available. This target is a TSH which is within the reference interval. If below it should be at least detectable (ie not suppressed below the limit of detection of the method). To achieve this, the FT4 will normally lie within or slightly above the reference interval. Different TSH methods have different limits of detection although there is inadequate evidence to discriminate between different levels of TSH suppression at these limits. We therefore use the term detectable by the method.

This strategy will prevent over-replacement in patients and decrease possible adverse effects noted in terms of cardiovascular outcome and loss of bone density. It has been suggested that in a minority of patients clinical well-being can only be achieved if the serum TSH is subnormal or suppressed and that this is of no detriment to the patient as long as the serum FT3 is unequivocally normal. However, no evidence could be found to support such a recommendation that in non-pregnant patients titrating the serum TSH to the lower half of the reference interval results in improved outcomes.

Direct effect on thyroid function (mostly suppression)

- amiodarone*
- lithium
- corticosteroids
- iodinated contrast media*
- other iodine preparations* (eg, over-the-counter kelp preparations)
- interferon a*
- dopamine, levodopa.

Analytical interference: increased FT4 from displacement

- heparin, via an increase in free fatty acids
- non-steroidal anti-inflammatories
- high-dose aspirin (.2 g/day).

Drugs increasing thyroxine replacement requirements

- c cytochrome P450 inducers: phenytoin, carbamazepine, ritonavir, rifampicin.

Intestinal absorbers

- sucralfate, colestyramine and colestipol, antacids containing aluminium
- ferrous sulphate
- proton pump inhibitors

Most commercial assays have now minimised assay interference per se, although users should refer to their local laboratory to discuss this possibility if unusual results are obtained.

*Iodine, amiodarone and interferon can produce either hypo- or hyperthyroidism, although the more common effect is hypothyroidism in western countries.

Box 1: Examples of drugs that influence thyroid measurements either through a pharmacodynamic effect on thyroid function or by binding displacement

Low TSH, normal FT4: does this patient have hyperthyroidism?

A GP receives the following thyroid function test result: FT4 16.2 pmol/l (reference interval 10-20 pmol/l) and TSH 0.1 mU/l (reference interval 0.4-4.5 mU/l) and contacts the laboratory for advice.

If the serum TSH is below the reference interval but above the limit of detection of the laboratory method:

- exclude non-thyroidal illness and drug effects (eg, patients on corticosteroid or dopamine therapy)
 - repeat 1 or 2 months later together with serum FT4 and FT3
 - continue to monitor if FT4/FT3 rising.

If the serum TSH is below the limit of detection of the laboratory method:

- measure serum FT4 and FT3 to exclude overt hyperthyroidism
- FT3 or FT4 above reference interval indicated hyperthyroidism.

If treatment is not undertaken, serum TSH should be measured every 6-12 months, and serum FT4 and FT3 measured if the serum TSH result is below 0.1 mU/l.

Patients with low TSH but normal FT4 and FT3 (free triiodothyronine) concentrations are deemed to have subclinical hyperthyroidism. Antibody testing is not recommended in these patients unless the clinical context is suggestive of hyperthyroidism, but those in which results are not explained by non-thyroidal illness or drug therapy should have tests repeated and if not treated, should be followed up with thyroid testing every 6-12 months. Endocrine referral is recommended for persistent subclinical hyperthyroidism. Free T3 measurement is require to identify cases of T3 toxicosis.

A drug history is important to identify patients taking pharmacological agents which may influence TSH or FT4 concentrations either because of their pharmacological activity (dopamine, corticosteroids, amiodarone) or assay interference (displaced protein binding with heparin, increasing FT4 concentrations).

Journal of Clinical Pathology, Volume 61(4), April 2008, pp 410-418

MICROBIOLOGY

Reappraisal of the Serum (1→3)-β-D-Glucan Assay for the Diagnosis of Invasive Fungal Infections-A Study Based on Autopsy Cases from 6 Years

Taminori Obayashi, Kumiko Negishi, Tomokazu Suzuki, and Nobuaki Funata

Background. The prevalence of invasive fungal infection is increasing. An effective diagnostic test is required to identify and treat them successfully.

Methods. All autopsy records at our hospital for the period from January 2000 through December 2004 were reviewed for cases of invasive fungal infection. The diagnostic efficacy of a serum (1→3)-β-d-glucan (β-glucan) assay was examined using only those cases in which patients had been tested for fungal infection within 2 weeks before death.

Results. Of 456 autopsies, 54 (11.8%) involved cases of invasive fungal infection. Leukemias were the most frequent underlying disease (in 52% of cases of invasive fungal infection), and *Aspergillus* species was the most frequent pathogen detected (in 70%). Of the 54 patients with invasive fungal infection, 41 had β-glucan testing performed within 2 weeks before death, as did 63 patients without invasive fungal infection; 48 of 54 patients with invasive fungal infection had a blood culture performed. The sensitivity and

specificity of the β -glucan test for the detection of invasive fungal infection were 95.1% and 85.7%, respectively, with a cutoff value of 30 pg/mL; 85.4% and 95.2%, respectively, with a cutoff value of 60 pg/mL; and 78.0% and 98.4%, respectively, with a cutoff value of 80 pg/mL. The sensitivity of blood culture testing was 8.3%. With a prevalence of 11.8%, the positive and negative predictive values for the β -glucan test were 47.1% and 99.2%, respectively, with a cutoff of 30 pg/mL; 70.4% and 98.0%, respectively, with a cutoff of 60 pg/mL; and 86.7% and 97.1%, respectively, with a cutoff of 80 pg/mL. During the 6-year period studied, of 21 patients with fungus-positive blood cultures that were preceded or followed by a β -glucan test within 2 weeks, 4 had negative β -glucan test results (β -glucan level, <30 pg/mL), and 17 had positive results (β -glucan level, >60 pg/mL); the concordance between culture results and β -glucan test results was 81.0%. Contrary to the general belief, 5 of 6 cases of cryptococemia were associated with high serum β -glucan levels.

Conclusion. The β -glucan test is an effective diagnostic tool for invasive fungal infection.

Clinical Infectious Diseases 2008;46:1864–1870