Frequency and Cause of Disagreements in Diagnoses for Fetuses Referred for Ventriculomegaly[1]

Deborah Levine, MD
Henry A. Feldman, PhD
João F. Kazan Tannus, MD, PhD
Judy A. Estroff, MD
Melissa Magnino, BA
Caroline D. Robson, MD
Tina Y. Poussaint, MD
Carol E. Barnewolt, MD
Tejas S. Mehta, MD, MPH
Richard L. Robertson, MD

Purpose: To prospectively assess the frequency and cause of disagreements in diagnoses at ultrasonography (US) and magnetic resonance (MR) imaging for fetuses referred for ventriculomegaly (VM).

Materials and Methods: One hundred ninety-five women, aged 18–44 years, with 200 fetal referrals for VM, were recruited in a prospective IRB-approved, HIPAA-compliant study. Written informed consent was obtained. US scans were prospectively interpreted by three obstetric radiologists and MR examinations were read by one obstetric radiologist and three pediatric neuroradiologists. Final diagnosis was reached by consensus (198 US, 198 MR, and 196 US-MR comparisons). Gestational age, ventricular size, types of disagreements, and reasons for disagreements were recorded. Interreader agreement was assessed with κ statistics. Ventricular diameter, gestational age, and confidence scores were analyzed by using mixed-model analysis of variance, accounting for correlation within reader and fetus.

Results: There was prospective agreement on 118 (60%) of 198 US and 104 (53%) of 198 MR readings. Consensus was more likely when the final diagnosis was isolated VM (83 of 104, 80% at US; 82 of 109, 75% at MR) than when the final diagnosis included other anomalies as well (14 of 63, 22% at US; seven of 68, 10% at MR; $P < .001$). There was disagreement on 19 (10%) of 196 and 31 (16%) of 196 fetuses about the presence of VM at US and MR, respectively, and on 29 (15%) of 198 and 39 (20%) of 198 fetuses regarding the presence of major findings at US and MR, respectively. Reasons for discrepancies in reporting major findings included errors of observation, lack of real-time US scanning, lack of neuroradiology experience, as well as modality differences in helping depict abnormalities.

Conclusion: Of radiologists who read high-risk obstetric US and fetal MR images for VM, there is considerable variability in central nervous system diagnosis.

[1] From the Department of Radiology, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215 (D.L., J.F.K.T., M.M., T.S.M.); and Clinical Research Program (H.A.F.) and Department of Radiology (J.A.E., C.D.R., T.Y.P., C.E.B., R.L.R.), Children's Hospital Boston, Boston, Mass. From the 2006 RSNA Annual Meeting. Received June 18, 2007; revision requested August 16; revision received August 29; accepted September 27; final version accepted October 12. Supported by National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering grant 01998. **Address correspondence to** D.L. (e-mail: *dlevine@caregroup.harvard.edu*).

There is a paucity of literature on the variability of readings in prenatal ultrasonography (US). While many articles have compared US with magnetic resonance (MR) imaging in prenatal central nervous system (CNS) diagnosis (1–7) and cited the need for high-quality US when a comparison with MR is made (1), in our opinion, the interobserver variability in US and MR diagnoses has not been adequately assessed. This is important, as prenatal imaging diagnosis is the basis for counseling the parents and, when postnatal imaging or autopsy is not available, is used for counseling purposes regarding recurrence risk in future pregnancies. Thus, our study was undertaken to prospectively assess the frequency and cause of disagreements in diagnoses at US and MR imaging for fetuses referred for ventriculomegaly (VM).

Materials and Methods

Patients and Imaging

Our study was performed at Beth Israel Deaconess Medical Center (Boston, Mass) and Children's Hospital (Boston, Mass) and was designed to compare outcomes of fetuses with VM. This study was funded by the National Institutes of Health, approved by the investigational review board, and compliant with the Health Insurance Portability and Accountability Act. Written informed consent was obtained. One hundred ninety-five women (age range, 18–44 years; mean, 31 years ± 5 [standard deviation]) with 199 fetuses (three sets of twins, each with VM; one woman who enrolled in the study each time for two pregnancies; and one woman who enrolled twice during the same pregnancy, once after dropping out of the study with initial examination findings interpreted as normal and again at a second examination with findings that were interpreted as abnormal) were involved in the study. These women were recruited between July 1, 2003, and August 20, 2006, with prenatal US findings that demonstrated VM (defined as ventricular size measured at the atrium of the lateral ventricle of ≥10 mm) or at US with a referral history of VM.

US and MR examinations were performed in a manner similar to that described by Levine et al (1) and evaluated as shown in Figure 1. Each fetus was referred for VM and each fetal examination was considered as a unique study. Date of last menstrual period, fetal age according to last menstrual period, and fetal biometry (including head circumference, biparietal diameter, and gestational age according to US) were recorded. Two patients with fetuses with neural tube defects (NTDs) underwent incomplete US prior to MR and thus were not included in the US analysis. Two patients withdrew from the MR portion of the study after undergoing US, thus they were not included in the MR analysis. Therefore, the study group for consensus opinions was 198 for US comparisons, 198 for MR comparisons, and 196 for combined US-MR comparisons. One hundred ninety-five (99.5%) of 196 US-MR examination sets were performed on the same day, one (0.5%) of 196 examination sets was performed within 1 day.

Confirmatory US Interpretation and Consensus

One of four obstetric radiologists (D.L., T.S.M., C.E.B., and J.A.E., with 7–20 years experience in high-risk obstetric US) measured the lateral ventricles at the level of the atria and coded anomalies by using a modification of the system delineated by Van der Knaap and Valk (8). The radiologist's confidence in diagnosis of the presence, character (appearance), and specific nature (specific diagnosis) of the abnormality was rated on a five-point scale (from a score of 1 = very confident, to a score of 5 = not confident).

US findings were then independently reinterpreted by the two obstetric radiologists who did not perform the US (to avoid any potential knowledge of the study prior to review), who also coded anomalies, lateral ventricle measurements, and confidence scales.

VM was diagnosed whenever one of the lateral ventricles at the level of the atrium was more than 10.0 mm in diameter. However, an obstetric radiologist could also diagnose VM subjectively,

Advances in Knowledge

- In fetuses referred for ventriculomegaly (VM), differences in opinion on the presence of central nervous system (CNS) anomalies are common, occurring in 40% of US studies and 47% of MR studies.
- Disagreements about the presence of major findings in association with VM are common, occurring in 15% of US studies and 20% of MR studies.
- Reasons for discrepancies in reporting of major findings at US and MR images when read by multiple readers include errors of observation, lack of real-time scanning, and lack of neuroradiology experience, as well as modality differences in helping depict abnormalities.

Implication for Patient Care

- Knowledge of reader variability in diagnosis of CNS anomalies is important when counseling patients carrying fetuses with VM since multiple imaging studies are often performed on these patients.

Published online before print
10.1148/radiol.2472071067

Radiology 2008; 247:516–527

Abbreviations:
ACC = agenesis of the corpus callosum
CNS = central nervous system
NTD = neural tube defect
VM = ventriculomegaly

Author contributions:
Guarantors of integrity of entire study, D.L., C.D.R.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, D.L.; clinical studies, D.L., J.F.K.T., J.A.E., C.D.R., T.Y.P., C.E.B., T.S.M., R.L.R.; statistical analysis, D.L., H.A.F.; and manuscript editing, D.L., H.A.F., J.A.E., T.Y.P., T.S.M., R.L.R.

Authors stated no financial relationship to disclose.

given the appearance of a dangling choroid plexus (9).

The US diagnoses of the three radiologists were compared (the reader performing US and the two from the other institution). Each US disagreement was coded as follows: *(a)* error of observation, *(b)* error of interpretation, *(c)* error of omission (a finding is clearly present but was not coded prospectively), *(d)* coding issue (two similar diagnoses with different codes, such as neuroproliferative porencephaly and encephaloclastic porencephaly), *(e)* disagreement regarding observation (still no agreement during consensus conference), *(f)* disagreement regarding interpretation (during consensus conference, the finding is seen by all but the interpretation of the finding was not agreed on), or *(g)* real-time scanning would have aided in diagnosis. In addition, final consensus was decided on by using a conference of three of the original readers, coded as being achieved by agreement of two or all three obstetric radiologists.

Disagreements (or consensus) in each fetus were coded as having complete agreement or one or more of the following: *(a)* decision to call VM (difference of opinion as to whether VM was present), *(b)* no clinical difference owing to disagreement (such as coding for Dandy Walker variant or inferior vermian hypoplasia or not coding a small midline cyst in a fetus with agenesis of the corpus callosum [ACC]), *(c)* minor new finding (for example, visualization of a portion of septal leaflets in a fetus that another reader coded as having complete absence of septal leaflets), *(d)* major new finding (such as an examination determined as showing findings of schizencephaly by one reader but not by another reader), *(e)* overcall of a minor finding (at consensus, the coded anomaly was determined as absent), or *(f)* overcall of a major finding. Examinations with disagreement on diagnosis were those coded as having new minor or major findings or those with overcalls of minor or major findings.

Examinations with and without diagnosis disagreement were compared with respect to gestational age.

MR Interpretation and Consensus

The MR examination was interpreted by one of the same group of radiologists who performed US (readers with 2–11 years experience with fetal MR), with knowledge of the findings of the confirmatory US. The readers coded the anomalies, measured the lateral ventricles at the level of the atria, and scored confidence in diagnosis. In addition, they coded for whether MR findings affected diagnosis by using the following scoring system: 1 = no effect, 3 = some effect, and 5 = critical information added.

Three pediatric neuroradiologists (C.D.R., T.Y.P., and R.L.R., each with 10–14 years experience in pediatric neuroradiology) then independently reviewed the MR examination with knowledge of referral diagnosis but without the aid of the confirmatory US and recorded their interpretation of the MR including ventricular size, coded CNS diagnosis, and confidence rating. The confirmatory US findings were then revealed and the neuroradiologist documented any change in diagnosis. Knowledge of the effect of the confirmatory US findings at final MR diagnosis was then scored on the same five-point scale as above.

MR consensus was reached at conference with the three neuroradiologists. MR disagreements were coded as follows: *(a)* error of observation, *(b)* error of interpretation, *(c)* error of omission, *(d)* coding issue, *(e)* disagreement regarding observation, or *(f)* disagreement regarding interpretation. In addition, final consensus was coded to denote agreement of two or all three neuroradiologists. Diagnosis disagreements were coded and compared with gestational age in a similar fashion to those described for US consensus.

US-MR Comparisons

The final US and MR diagnoses were compared by one of the authors (D.L.), in consultation with the neuroradiologists (C.R., R.L.R., T.Y.P.), when any final diagnosis was unclear. Disagreements were coded as follows: *(a)* disagreement regarding observation, *(b)* disagreement regarding interpretation, *(c)* error of omission, *(d)* coding issue, *(e)* finding expected to be better seen at US (such as the wall of an arachnoid cyst), *(f)* finding expected to be better seen at MR (such as migrational abnormality), or *(g)* neuroradiologist experience would aid in diagnosis.

Figure 1

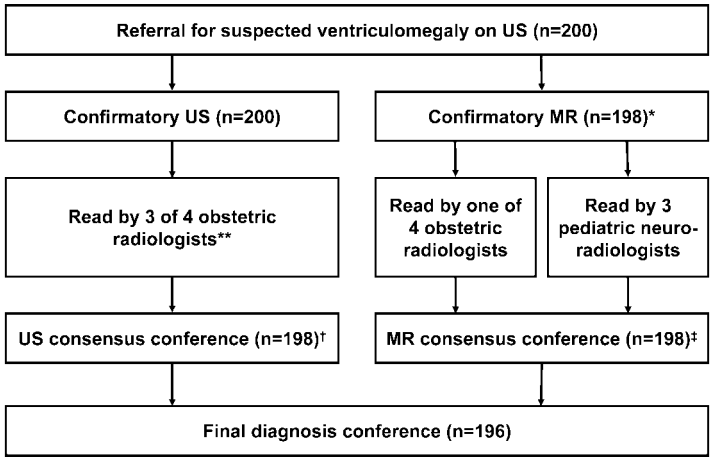


Figure 1: Study design. * = MR not performed in two fetuses. ** = US performed by one obstetric radiologist and independently read by two obstetric radiologists from second institution. † = Two fetal US with NTD findings were incomplete at MR and not included in US consensus. ‡ = MR consensus performed by three pediatric neuroradiologists used to assess need for consensus diagnosis by obstetric radiologist who read MR, diagnoses of neuroradiologists.

Statistical Analysis

Interreader agreement was assessed with κ statistics. Ventricular diameter, gestational age, and confidence scores were analyzed by using mixed-model analysis of variance, accounting for correlation within reader and fetus (Appendix).

Results

Fetal age measured by using date criteria had a mean of 26.2 weeks ± 5.8 (range, 16.3–41.0 weeks). Fetal age measured by using US had a mean of 26.3 weeks ± 5.9 (range, 15.7–39.4 weeks). Of the diagnoses (Table 1), the most frequently occurring were VM, dysgenesis of the corpus callosum, and normal, and showed good to excellent agreement among US and MR readers, as indicated by the κ statistic (Table 2). Chiari malformation and spinal NTD showed excellent agreement.

US Diagnoses and Agreement

Of 198 US studies, there were 118 (60%) in which the three US readers agreed prospectively (preconference) on all diagnoses. These final diagnoses included isolated VM (n = 83), normal (n = 19), holoprosencephaly (n = 2), VM and other anomalies (n = 14), ACC (n = 7, one with midline cyst), spinal NTD (n = 4), choroid plexus cyst (n = 2), and Dandy Walker syndrome (n = 1). The preconference consensus was more likely when the final diagnosis was isolated VM (83 of 104, 80%) than VM and anomalies (14 of 63, 22%; $P <$ .001) (Fig 2).

There was disagreement in 19 (10%) of 196 studies regarding the presence of VM at US. Ventricular diameter varied significantly according to the level of agreement among the three US readers concerning presence of VM (Table 3). Studies with preconference agreement on the presence of VM were significantly larger (mean, 13.5 mm) than were those with preconference agreement on the absence of VM (mean, 7.5 mm). In studies with disagreement, regardless of the ultimate consensus, those readers who indicated VM was present consistently recorded larger ventricular diameter for a given fetus (mean, 10.2 mm) than those who indicated VM was absent (mean, 8.4 mm).

Variation among the four obstetric radiologists was 0.3 mm in ventricular diameter for a given fetus, compared with 4.5 mm among fetuses and a 1.3-mm residual error. The intraclass correlation (the fraction of variance attributable to actual differences among cases) was 92%, whereas interreader variability represented less than 1% of the total variance.

There were 80 (40%) of 198 studies without consensus on the prospective US readings (Table 4), with 153 disagreements on specific diagnoses(Table 5). Consensus was achieved with agreement of all three obstetric radiologists for 122 (80%) of these 153 diagnoses and with agreement of two for the remaining 31 (20%) diagnoses. Twenty-seven disagreements were categorized as having no clinical implication. Of the 80 studies requiring a conference (Table 4), there was disagreement regarding the presence of VM in 19 fetuses and in 29 (15%) regarding the presence of major findings.

Coding issues (50 of 153, 33%) and errors of observation (51 of 153, 33%) were the most common types of disagreement (Table 5). In only five (3%) was it thought that real-time scanning would have aided in diagnosis.

Table 1

Diagnoses Made by Using US and MR in 196 Examinations

Diagnosis	No. of US Examinations*		No. of MR Examinations*		
	Performing Imager	Consensus	Performing Imager	Consensus	Final Consensus
Normal	20 (3)	23 (2)	19 (5)	15 (3)	23
VM	165 (5)	166	161 (6)	165 (10)	168
Dysgenesis of corpus callosum†	24 (3)	22 (1)	23 (3)	27 (1)	27
Spinal NTD	9	9	10	10	10
Hemorrhage	7	7	8	11	12
Porencephaly‡	6 (2)	6 (3)	9 (2)	9	10
Defect of septi pellucidi	4	5	5	9 (1)	9
Chiari malformation	8	9	9	8	9
Cerebellar hypoplasia	7 (3)	9	7 (2)	8	10
Cyst§	7	10	7	6	12
Polymicrogyria, lissencephaly	4	3 (1)	7 (3)	9	9
Dandy Walker variant and/or malformation	4 (2)	4 (2)	4 (1)	5	5
Congenital infarction	1	0	2	5	5
Holoprosencephaly	2	2	2	3	3
Schizencephaly	0	0	2	2	2
Megacisterna magna	2 (1)	2 (1)	2	2	3
Abnormal myelination	0	0	0	1	1
Kinked midbrain	0	1	1	1	1
Encephalocele	0	0	1	1	1
Craniosynostosis	0 (1)	0	0	1	1
Periventricular leukomalacia	0	0	0	1	1
Micrencephaly	1 (1)	0	1	1	1
Tumor	1	1	1	1	1
Heterotopias	1	1	0	1	1

* Numbers in parentheses are overcalls (diagnoses deemed to be not present at final consensus conference).

† In one of these studies overcalled at US and MR, the neuroradiologist's correlative MR diagnosis was holoprosencephaly.

‡ In the two examinations overcalled at US and MR, the neuroradiologist's correlative MR diagnoses were schizencephaly in one case and abnormal myelination in the other.

§ Includes two choroid plexus cysts.

Table 2

Agreement between US and MR Readers

Diagnosis	κ Value for Examinations				
	US Only	Neuroradiologist MR Only	All MR	All US and MR	US Consensus vs MR Consensus
Normal	0.79	0.71	0.75	0.74	0.66
VM	0.74	0.63	0.68	0.66	0.66
Dysgenesis of corpus callosum	0.84	0.70	0.75	0.78	0.84
Spinal NTD	1.00	0.90	0.93	0.94	0.95
Hemorrhage	0.67	0.65	0.67	0.66	0.66
Porencephaly	0.43	0.53	0.59	0.52	0.55
Defect of septi pellucidi	0.59	0.56	0.56	0.52	0.67
Chiari malformation	0.86	0.97	0.97	0.91	0.94
Cerebellar hypoplasia	0.33	0.21	0.28	0.34	0.81
Cyst	0.66	0.45	0.53	0.55	0.47
Polymicrogyria, lissencephaly	0.26	0.48	0.52	0.37	0.44
Dandy Walker variant and/or malformation	0.69	0.61	0.69	0.66	0.72
Standard error*	0.04	0.04	0.03	0.02	0.07

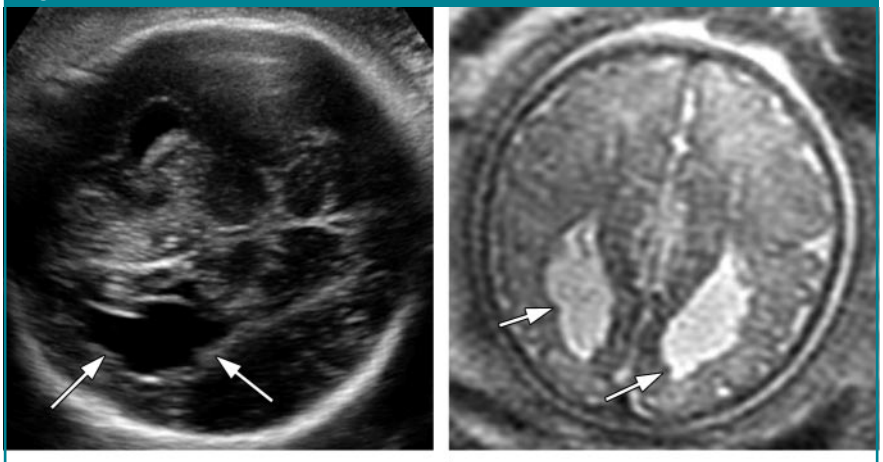
* Standard error for multireader κ depends only on sample size parameters and applies to all diagnoses (10).

The level of consensus of the US readers concerning presence of VM was not affected by head circumference, biparietal diameter, or gestational age ($P > .40$) (Table 6). There was a significant difference among these groups when comparing ventricular diameter: those fetuses agreed on as having VM, with the largest ventricles (mean, 13.5; range, 8–44 mm); those agreed on as not having VM, with the smallest ventricles (mean, 7.5; range, 4–9 mm); and those with disagreement regarding VM, with ventricles of intermediate size (mean, 9.4; range, 7–15 mm).

Confidence regarding the presence of a CNS abnormality was not affected by agreement being present among the US readers (Fig 3), whereas confidence in the character of abnormality, type of abnormality, and presence of additional findings were all significantly higher when the US readers agreed.

There was no relationship between gestational age and diagnosis disagreement at US.

Figure 2



a. b.

Figure 2: ACC with heterotopias at 35 weeks gestational age. **(a)** Transverse US shows colpocephaly with slitlike frontal horns and areas of increased echogenicity (arrows) lining ventricles. **(b)** Transverse T2-weighted MR image (echo spacing, 4.2 msec; echo time, 60 msec; echo train length, 72; one acquisition; section thickness, 4 mm; field of view, 26 × 30 cm; matrix, 192 × 256; sequence acquisition, 16 seconds; and section acquisition, 420 msec) shows similar findings with areas of low signal intensity projecting into ventricle (arrows). ACC was coded by all readers, but heterotopias were coded by only one of three US readers and all three neuororadiologists.

MR Diagnoses and Agreement

Of 198 MR studies, the four readers reached consensus prospectively in 104 (53%). These final diagnoses included isolated VM ($n = 82$), normal ($n = 14$), holoprosencephaly ($n = 1$), and VM and another diagnosis ($n = 7$; four with ACC, two with spinal NTD, and one with CNS neoplasm). Consensus was significantly more likely when VM was isolated (82 of 109, 75%) than when other anomalies were present in association with VM (seven of 68, 10%; $P < .001$)

Excluding two fetuses with holoprosencephaly, there was disagreement about the presence of VM at MR

in 31 (16%) of 196 fetuses. Ventricular diameter was largest in fetuses with preconference agreement on VM (mean, 14.4 mm) and smallest in fetuses with preconference agreement that VM was absent (mean, 7.4 mm; $P < .001$) (Table 3). In studies with disagreement, readers favoring VM recorded larger ventricular diameter for a given fetus (10.9 vs 9.0 mm, $P < .001$). Variation among the readers for a given fetus and reader classification (obstetric radiologist or neuroradiologist) was 0.4 mm, representing less than 1% of total variance. The intraclass correlation was 93%.

There were 94 (47%) of 198 studies without consensus on the prospective MR readings (Table 4), with 223 disagreements on particular diagnoses (Table 5). The three neuroradiologists reached unanimous consensus in conference on 218 (98%) of those 223 disagreements and resolved the remaining five (2%) with a two-to-three consensus. Of the 77 diagnoses where the disagreement included something other than the decision to call VM, 32 concerned coding issues of no clinical consequence. There were 39 (20%) fetuses with disagreement regarding the presence of major findings (Table 4).

Coding issues (82 of 223, 37%) and errors of observation (86 of 223, 39%) were most common types of disagreement (Table 5). Among the MR readers, head circumference did not vary according to the level of consensus concerning presence of VM ($P = .24$) but biparietal diameter was larger (mean, 10 mm) and gestational age was older (mean, 3.3 weeks) in fetuses where the ultimate MR consensus included VM (Table 6).

MR readers' confidence in the presence, character, or type of abnormality did not differ between examinations with agreement and disagreement (Fig 3), but their confidence in additional findings was significantly higher in examinations where they agreed ($P < .03$).

In five (3%) of 198 studies where neuroradiologists interpreted the images before and after use of US findings, the knowledge from the US findings changed the MR interpretation, including one fetus with intraventricular hemorrhage and porencephaly (seen on only three MR images in the study); and one fetus each with small subdural hematoma, megacisterna magna, cerebellar hypoplasia, and callosal dysgenesis. In one study, this was scored as having no effect on reading, in two studies as having some effect, and in three studies as having a critical effect on reading.

There was no relationship between gestational age and diagnosis disagreement at MR.

Table 3

Ventricular Diameter According to Modality and According to Preconference Agreement for Presence of VM

Modality and Agreement*	No. of Examinations†	No. of Measurements	Ventricular Diameter (mm)‡
US			
All	196	588	12.5 ± 0.2
Readers agree, VM present	157	471	13.5 ± 0.4
Readers agree, VM absent	20	60	7.5 ± 1.0
Readers disagree	19	57	9.4 ± 1.1
Reader indicates VM present	. . .	31	10.2 ± 1.1
Reader indicates VM absent	. . .	26	8.4 ± 1.1
Difference			1.7 ± 0.4
MR			
All	196	784	13.1 ± 0.2
Readers agree, VM present	150	600	14.4 ± 0.4
Readers agree, VM absent	15	60	7.4 ± 1.2
Readers disagree	31	124	10.2 ± 0.9
Reader indicates VM present	. . .	77	10.9 ± 0.9
Reader indicates VM absent	. . .	47	9.0 ± 0.9
Difference			2.0 ± 0.3
US-MR comparison			
All	194	1358	12.8 ± 0.1
US-MR consensus, VM present	164	1148	13.5 ± 0.4
US-MR consensus, VM absent	18	133	7.6 ± 1.1
US-MR consensus disagree	12	84	9.7 ± 1.3
Reader indicates VM present	. . .	42	10.8 ± 1.3
Reader indicates VM absent	. . .	42	8.5 ± 1.3
Difference			2.3 ± 0.3

* Cases of agreement include those where readers disagreed regarding diagnoses other than VM or agreed that the diagnosis made no clinical difference.

† Number of fetuses, each measured by three US or four MR readers. Two fetuses with holoprosencephaly were excluded.

‡ Data expressed as mean or difference ± standard error and measured by using analysis of variance, adjusted for within-case and within-reader correlation. All means differ significantly according to readers' agreement, $P < .001$. MR results also adjusted for reader class (obstetric radiologist or neuroradiologist). US-MR results also adjusted for reader class and modality.

Comparing US and MR Readings

Of 147 US-MR pairs of examinations, the obstetric radiologist who performed and interpreted US and MR images was the same, including 99 (99%) of 100 from one institution and 46 (48%) of 96 from the second. There were five fetuses with diagnosis disagreement, with the final diagnosis being visualized on US and not MR, including two fetuses with cysts associated with ACC and one fetus each with ACC, porencephaly, and cerebellar hypoplasia with heterotopias (visualized at MR, as coded by neuroradiologists). There were 14 pregnancies with diagnosis disagreement with final diagnosis seen at MR and not at US when read by the same reader, including four fetuses with porencephaly; three fetuses with polymicrogyria; two fetuses with schizencephaly; and one fetus each with subependymal hemorrhage, spinal meningocele, encephalo-

Table 4

Disagreement Categories According to Case

Category of Disagreement*	No. of US Cases (n = 80)		No. of MR Cases (n = 94)		No. of US-MR Cases (n = 62)	
	Any	Sole†	Any	Sole†	Any	Sole†
Decision to call VM	19 (24)	14 (18)	32 (34)	17 (18)	13 (21)	12 (19)
No clinical difference due to disagreement	27 (34)	19 (24)	32 (34)	20 (21)	18 (29)	17 (27)
Decision to call VM or no clinical difference owing to disagreement	42 (53)	37 (46)	57 (61)	42 (45)	30 (48)	30 (48)
New major finding	12 (15)	11 (14)	18 (19)	8 (9)	21 (34)	20 (32)
New minor finding	12 (15)	8 (10)	14 (15)	5 (5)	6 (10)	6 (10)
Overcall major finding	17 (21)	12 (15)	21 (22)	9 (10)	4 (6)	3 (5)
Overcall minor finding	6 (8)	3 (4)	8 (9)	8 (9)	2 (3)	2 (3)

Note.—Numbers in parentheses are percentages.

* Number of fetuses in indicated category among all cases of disagreement. Categories are not mutually exclusive.

† The only disagreement recorded in the fetuses.

Table 5

Types of Disagreement According to Diagnosis

Type of Disagreement	No. of US Cases (n = 153)		No. of MR Cases (n = 223)		No. of US-MR Cases (n = 103)	
	Any	Sole*	Any	Sole*	Any	Sole*
Error of observation	50 (33)	47 (31)	86 (39)	63 (28)	12 (12)	10 (10)
Error of interpretation	13 (8)	13 (8)	61 (27)	43 (19)	14 (14)	9 (9)
Error of omission	7 (5)	7 (5)	12 (5)	11 (5)	3 (3)	3 (3)
Coding issue	51 (33)	51 (33)	82 (37)	65 (29)	23 (22)	21 (20)
Disagreement regarding observation	14 (9)	12 (8)	9 (4)	6 (3)	. . .	. . .
Disagreement regarding interpretation	17 (11)	17 (11)	5 (2)	4 (2)	. . .	. . .
Real-time scanning would have helped	5 (3)	5 (3)	. . .	. . .	. . .	. . .
Neuroradiology experience would have helped	. . .	. . .	. . .	. . .	15 (15)	11 (11)
Expect to see better at US	. . .	. . .	. . .	. . .	8 (8)	8 (8)
Expect to see better at MR	. . .	. . .	. . .	. . .	22 (21)	21 (20)

Note.—Number of diagnoses in indicated category, among all diagnoses on which prospective disagreement occurred. Types are not mutually exclusive. Numbers in parentheses are percentages.

* The only disagreement recorded in the fetuses.

cele, congenital infarction, and septooptic dysplasia.

When comparing ventricular measurements from all seven readers, three for US and four for MR (Table 3), the recorded diameter was largest in fetuses for which US and MR consensuses included VM (mean, 13.5 mm) and smallest in fetuses where US and MR consensuses excluded VM (mean, 7.6 mm; $P < .001$). In studies where US and MR consensuses did not match, those readers favoring VM recorded larger ventricular diameter for a given fetus by a margin of over 2 mm in the mean diameter (10.8 vs 8.5 mm, $P < .001$). Variation among the readers for a given fetus and reader class was 0.4 mm, representing less than 1% of total variance. The intraclass correlation was 92%. MR measurements of ventricular diameter were greater than those from US by 0.6 mm on average (standard error, 0.3 mm; $P = .05$) for a given fetus, all other factors being equal (reader class and degree of agreement).

Of the 196 US-MR cases, there were 83 (42%) diagnoses with complete agreement by the seven readers. These diagnoses included normal ($n = 11$), isolated VM ($n = 68$), VM with NTD ($n = 1$), and VM with ACC ($n = 3$).

The consensus US opinion and the consensus MR opinion were in agreement on all diagnoses in 134 (68%) of 196 fetuses. The 62 (32%) fetuses with disagreements included 30 where the only disagreements were either to diagnose VM or were differences in coding judged to have no clinical effect (Table 4). The remaining 32 (16%) with diagnosis disagreement included 21 major new findings, six minor changes in diagnosis, four overcalls of major findings, and two overcalls of a minor finding. The 21 major new findings included fetuses with meningocele, porencephaly, hemorrhage, schizencephaly, and other migrational abnormalities (Table 1). The overcalls of major findings included a fetus at 22 weeks with an MR-aided diagnosis of ACC not agreed on at consensus, a

fetus at 21 weeks with VM and ACC and a question of encephaloclastic porencephaly that was not agreed on at consensus, and a fetus at 35 weeks with VM and porencephaly with a question of Dandy Walker variant at US where the Dandy Walker variant was not agreed on at consensus.

Head size and gestational age were not associated with the US-MR consensus (Table 6).

The four radiologists performing MR varied considerably in their ratings of the effect of MR on diagnostic certainty beyond what US alone would provide (Table 7). Two obstetric radiologists had lower confidence scores at US than did the other two; those with the lowest confidence score at US rated the highest effect of MR. Adjusting for these individual tendencies, the effect of MR was significantly greater in examinations where the performing reader disagreed with the neuroradiologists on the diagnosis of VM by 0.4–0.8 points on a scale of 1–5, than in examinations where they agreed or judged their difference to be clinically unimportant.

When the same readers examined US and MR readings, they recorded virtually identical levels of confidence in the presence, character, and type of abnormality ($P > .5$) but were slightly more confident about additional MR findings associated with the abnormality (0.7 points, $P = .004$). The readers indicated that MR provided somewhat more diagnostic certainty in examinations where they had less confidence in the US diagnosis; this correlation was mild (Spearman $r = 0.2–0.5$) but significantly different from zero.

There was no relationship between gestational age and diagnosis disagreement on US-MR comparisons.

Discussion

It is well recognized that measurement variability can be a factor in prenatal diagnosis (10,11). Measurement of the lateral ventricle is subject to errors owing to an off-axis image plane of a section, an angled measurement, or improper choice of ventricular boundary leading to a relatively large number of false-positive test results (12). Differences in opinion in the diagnosis of VM were prevalent in our population and may have resulted, in part, from these factors. When the ven-

Table 6

Measures of Head Size and Gestational Age as Related to Reader Consensus Concerning Presence of VM

Consensus on Modality, VM Presence, and Preconference Agreement	No. of Cases*	Biparietal Diameter (mm)†	Head Circumference (mm)†	Gestational Age According to Date (wk)†	Gestational Age by Using US (wk)‡
US					
Present					
Agree	157	63 ± 2	231 ± 5	26.1 ± 0.5	26.1 ± 0.5
Disagree	10	71 ± 6	271 ± 21	29.1 ± 2.0	28.5 ± 1.8
Absent					
Agree	20	60 ± 4	230 ± 14	25.2 ± 1.3	24.9 ± 1.3
Disagree	9	69 ± 6	245 ± 20	27.8 ± 2.1	27.5 ± 1.9
P value		.60	.42	.49	.48
MR					
Present					
Agree	150	65 ± 2	237 ± 5	26.7 ± 0.5	26.5 ± 0.5
Disagree	26	64 ± 4	224 ± 12	25.7 ± 1.2	26.0 ± 1.1
Absent					
Agree	15	55 ± 5	212 ± 16	23.4 ± 1.5	23.6 ± 1.5
Disagree	5	51 ± 9	201 ± 35	22.2 ± 2.9	21.6 ± 2.6
P value		.04	.24	.05	.02
US-MR comparison					
Present					
Agree	164	64±2	233 ± 5	26.3 ± 0.5	26.2 ± 0.4
Disagree	4	58 ± 10	220 ± 35	23.7 ± 2.9	23.6 ± 2.9
Absent					
Agree	19	58 ± 4	221 ± 15	24.3 ± 1.4	23.9 ± 1.4
Disagree	9	73 ± 6	260 ± 22	29.8 ± 2.2	29.3 ± 2.0
P value		.46	.54	.31	.39

Note.—US measurements and gestational age scored by a single reader (one of four obstetric radiologists). Mean ± standard error from mixed-model analysis of variance, adjusted for within-reader correlation. *P* value tests for equal mean in cases of VM presence.

* There were 196 US, 196 MR, and 194 US-MR cases; two fetuses with holoprosencephaly were excluded.

† Obtained by using US.

‡ According to last menstrual period.

tricle measures close to 10 mm, some radiologists tended to diagnose VM and others tended to call the fetus normal. In our series of 200 fetal examinations referred for VM, the ultimate consensus was that 168 (84%) had VM. There were 19 (10%) of 198 and 31 (16%) of 198 fetuses with disagreement about the presence of VM at US and MR, respectively. This reinforces the importance of standardization of the measurement plane.

In addition, measurements obtained with MR were larger than those obtained with US, which can also affect the rate of diagnosis of VM. It is important to acknowledge that these measurement differences occur, since patients often have more than one imaging study in the work-up of VM. While the degree of VM may change over time, it is also possible that some differences in diagnosis result from individual variability in measurements and modality. It may be that fetuses diagnosed with mild VM in whom reviewers cannot agree if VM is present will have better outcomes than fetuses with consensus about the diagnosis of VM.

US and MR are both used to characterize fetal CNS abnormalities; each modality has specific strengths and weaknesses in helping depict abnormalities. For example, US may show the wall of an arachnoid cyst to better advantage than does MR, whereas MR frequently shows cortical migrational abnormalities to better advantage than does US. While these generalities are well recognized, the variability in individual interpretation of studies has not, to our knowledge, been assessed.

Diagnostic image interpretation is dependent on detection and characterization of findings. US-aided diagnosis is known to depend on the skill of the person obtaining the images to correctly display the anatomy and the skill of the reader to appropriately diagnose the abnormality. When assessing US, we are limited by the images that have been obtained, with real-time scanning at times aiding in diagnosis. The Routine Antenatal Diagnostic Imaging with US study demonstrated the importance of user experience in fetal anomaly detec-

Figure 3

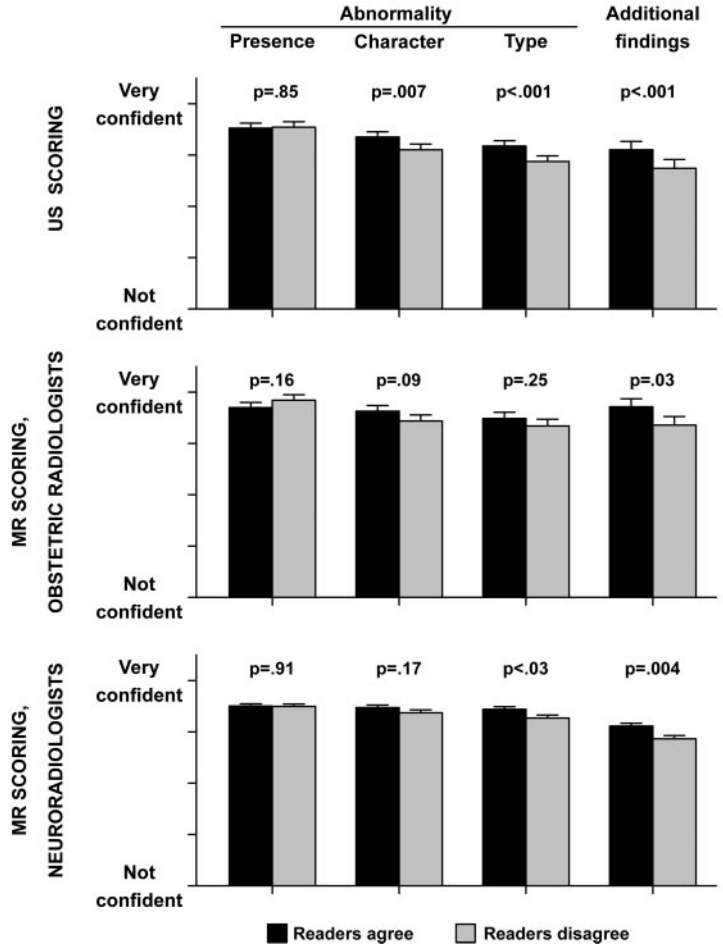


Figure 3: Confidence of US and MR readers in diagnosis of fetal CNS abnormalities. Bars indicate mean ± standard error on five-point scale from very confident to not confident. *P* values from analysis of variance are adjusted for interreader and intersubject variability.

Table 7

Effect of MR on Diagnostic Certainty for CNS Abnormalities

Reader	Cases	Presence*	Character*	Type*	Associated Findings*
Readers agree†	87	2.1 ± 0.2	2.1 ± 0.2	2.0 ± 0.2	1.5 ± 0.2
Readers disagree	74	2.5 ± 0.2	2.9 ± 0.2	2.8 ± 0.2	2.3 ± 0.2
P value	. . .	.01	<.001	<.001	<.001
Reader A	41	4.1 ± 0.2	4.4 ± 0.2	4.1 ± 0.2	3.0 ± 0.2
Reader B	57	2.6 ± 0.2	2.8 ± 0.2	2.5 ± 0.2	1.8 ± 0.3
Reader C	95	1.5 ± 0.1	1.9 ± 0.1	1.7 ± 0.1	1.8 ± 0.1
Reader D	5	1.0 ± 0.5	1.1 ± 0.5	1.5 ± 0.5	1.1 ± 0.6

* Data are the mean ± standard error (except for *P* values) on a five-point scale (from 1 = no improvement, to 5 = significant improvement); measured by using analysis of variance adjusted for interreader differences. *P* values test for difference between cases of reader agreement and disagreement.

† Obstetric radiologist obtaining images agrees prospectively with three neuroradiologists on presence of VM, including cases where disagreement lacks clinical consequence.

tion, since highly experienced ultrasonographers detected 35% of anomalies compared with only 13% by less-experienced US readers (13). In our study, we demonstrated that even individuals with expertise in obstetric US, when evaluating the same images, can arrive at different fetal diagnoses.

When assessing CNS anomalies, another factor to consider is the subspecialty training of the individual interpreting the findings of the examination. The pediatric neuroradiologists in our study diagnosed more abnormalities when interpreting fetal MR than did the obstetric radiologists. It is likely that some of the perceived benefit of MR in previously published studies has come from the effect of having neuroradiology input with the final diagnosis.

Another factor to consider in US-aided diagnosis of fetal CNS anomalies is experience with visualization of the specific anomalies. We have found improved recognition of such diagnoses as ACC and heterotopias at US in our study compared with prior studies (14,15), likely owing to improved understanding of how these anomalies appear in utero.

Prior studies that compared US with MR in aiding the diagnosis of fetal brain abnormalities stated the importance of having high-quality US images available for comparison to assess the incremental benefit of MR (1). However, in clinical practice this may not be as important a factor. In our study of a highly enriched population of fetal CNS abnormalities, knowledge of the confirmatory US findings was only judged to have affected diagnosis in 2.5% of examinations. This is a reassuring finding for clinical use of fetal MR because in many centers, the imaging expertise with obstetric US and fetal MR do not coexist in the same radiology department.

We found 29 (14.8%) of 196 US and 39 (19.9%) of 196 MR examinations had disagreements about major findings being present. Our disagreements and types of errors are similar to those originally categorized by Smith (16) in 1967 and updated by Renfrew et al (17) These studies show false-positive errors of overreading and misinterpretation, true-positive readings with misclassification, and false-negative errors. We found false-negative diagnoses in 33% and 39% of US and MR disagreements, respectively.

An example of a false-negative finding that can be attributed to satisfaction of search was in a fetus with VM with a small meningocele without Chiari malformation. It is likely that a more extensive US search of the spine and cranium would have depicted this small lesion. An example of false-negative findings at MR is the difficulty in navigation through large data sets where error can occur, as in a fetus with porencephaly in the frontal horn that was noted by one neuroradiologist only after confirmatory US results were given, indicating the region of abnormality.

Other reasons for false-negative findings were lack of real-time scanning for US diagnoses in 3% of US disagreements and lack of neuroradiology training for US-MR comparisons in 15% of MR disagreements.

After consensus, we found that MR added additional diagnoses beyond those found with US. In particular, fetuses with migrational abnormalities and porencephaly were diagnosed at MR but not at US. These findings are similar to findings in other studies that have assessed the effect of MR on prenatal diagnosis of CNS anomalies (1,18–23). This reinforces the added value of MR in fetuses referred for VM. Another factor that affects the perceived benefit of MR is the confidence in diagnosis of the radiologist. In our study, two of the obstetric radiologists had lower confidence scores than did the other two. The individual with the lowest confidence score for US rated the highest effect for MR.

A limitation of our study was the lack of a true reference standard. We used the opinion of the US-MR consensus conference as our final diagnosis. Because our study assessed the reproducibility of reporting of individual findings rather than the sensitivity of detecting a final outcome at birth, we did not feel that postnatal outcome was needed. This reference standard of imaging concordance has been used by others (24–26).

A second limitation was that the study situation may have led the radiologists to identify more subtle abnormalities than would otherwise be noted. However, in a clinical setting of a high-risk population of fetuses referred for VM, the threshold level for suggesting abnormality is likely similar. The alternative is also possible, in that the effects of the study situation might have led second-opinion readers to take less care than usual, knowing that their impressions would not be used for immediate patient care.

A third limitation was that by coding abnormalities, there was no "gray zone." Radiologists had to decide if an anomaly was present, with no option to indicate uncertainty in each finding. This was necessary to compare the readings of the individuals but probably led to an overestimate of discrepancies.

A fourth limitation was that the results are only applicable to fetuses with VM. Our patient population does not reflect that which would be found in a screening population or among those referred for MR for any other indication. In a screening population, we would expect much higher concordance in findings, since a large percentage of the fetuses would be normal.

In conclusion, among radiologists who read high-risk obstetric US and fetal MR for VM, there is considerable variability in CNS diagnoses. A reduction in variability will require more consistent criteria for diagnostic interpretation. We assessed some of the reasons for missing and misdiagnosing some of the anomalies. Knowledge of the appearance of anomalies seen in association with VM should help to improve the accuracy of prenatal diagnosis.

Appendix

Strength of agreement among readers was evaluated with respect to normal, VM, and other diagnoses by using the κ statistic, a dimensionless index with κ = 0–0.20 indicating poor agreement, κ = 0.21–0.40 indicating fair agreement, κ = 0.41–0.60 indicating moderate agreement, κ = 0.61–0.80 indicating good agreement, and κ = 0.81–1.00 in-

dicating excellent agreement (27,28). κ Statistics were separately calculated for the most common diagnoses for the following groups: the group of four obstetric radiologists (three of whom read a given scan); the group of three pediatric neuroradiologists (all of whom read each MR); the group of four readers who participated in MR diagnoses (one obstetric radiologist and three neuroradiologists for each image); and the entire group of seven US and MR readers and the consensus US and MR readings.

The association between head circumference, biparietal diameter, and gestational age and diagnostic agreement concerning VM was evaluated by dividing the cases into four classes according to the degree of agreement achieved in subsequent stages of the study: *(a)* preconference consensus of VM, *(b)* preconference consensus on absence of VM, *(c)* initial disagreement resolved as a final diagnosis of VM, and *(d)* initial disagreement resolved as absence of VM. This classification was assigned separately for each stage of the study (US, MR, and US-MR). Comparison across the classes was made by using analysis of variance. To account for within-reader correlation, the analysis of variance included a random effect identifying the obstetric radiologist who performed the measurement.

Ventricular diameter was measured by all readers, permitting us to examine the association between individual reader measurements of ventricular size with respect to fetal head size and/or gestational age and the subsequent degree of disagreement concerning presence of VM. We divided the cases into three classes: *(a)* all readers at preconference who agreed on presence of VM, *(b)* all readers at preconference who agreed on absence of VM, and *(c)* readers who disagreed regarding VM. Measurements for the third class were further subdivided as those from individual readers who scored VM as either present or absent.

The entire set of measurements (three per fetus for US, four per fetus for MR, and seven per fetus for US-MR) was subjected to analysis of variance, with the class of agreement as independent variable. Analyses that included MR measurements were adjusted for a dichotomous indicator of whether the reader was an obstetric radiologist or a neuroradiologist. Analysis of the combined MR-US measurements was additionally adjusted for modality, enabling us to compare US with MR, controlling for all other factors. To account for variability among fetuses and systematic variability among individual readers, the analysis of variance included random effects. Intraclass correlation, defined as the fraction of variance attributable to true variability in the population (as opposed to measurement artifact), was calculated as interfetus variance divided by the sum of interfetus, interreader, and residual variance.

In two fetuses that were coded as holoprosencephaly by all readers, lateral ventricles were not measured by some readers, and VM was not coded by some readers. These cases were excluded from the above analyses of ventricular size.

The effect of MR on visualization of anomalies was assessed by comparing scores from readers who read US and MR images, and by comparing the consensus diagnoses reached at US and MR. Confidence scores were compared for fetuses with and without diagnosis disagreement. Analysis of variance was employed for the analyses of confidence and effect, with reader effects assessed from fitted parameters of the model.

The association between gestational age and diagnosis disagreement was performed with the Student *t* test.

A *P* value of less than .05 was considered to indicate a significant difference. All computations were performed with software (SAS, version 9.1; SAS Institute, Cary, NC).

References

1. Levine D, Barnes PD, Robertson RR, Wong G, Mehta TS. Fast MR imaging of fetal central nervous system abnormalities. Radiology 2003;229:51–61.
2. Levine D, Barnes PD, Madsen JR, Abbott J, Mehta T, Edelman RR. Central nervous system abnormalities assessed with prenatal magnetic resonance imaging. Obstet Gynecol 1999;94:1011–1019.
3. Poutamo J, Vanninen R, Partanen K, Ryynanen, Kirkinen P. Magnetic resonance imaging supplements ultrasonographic imaging of the posterior fossa, pharynx and neck in malformed fetuses. Ultrasound Obstet Gynecol 1999;13:327–334.
4. Dinh DH, Wright RM, Hanigan WC. The use of magnetic resonance imaging for the diagnosis of fetal intracranial anomalies. Childs Nerv Syst 1990;6:212–215.
5. Levine D, Barnes PD, Madsen JR, Li W, Edelman RR. Fetal central nervous system anomalies: MR imaging augments sonographic diagnosis. Radiology 1997;204:635–642.
6. Twickler DM, Magee KP, Caire J, Zaretsky M, Fleckenstein JL, Ramus RM. Second-opinion magnetic resonance imaging for suspected fetal central nervous system abnormalities. Am J Obstet Gynecol 2003;188:492–496.
7. Hubbard AM. Ultrafast fetal MRI and prenatal diagnosis. Semin Pediatr Surg 2003;12:143–153.
8. van der Knaap MS, Valk J. Classification of congenital abnormalities of the CNS. AJNR Am J Neuroradiol 1988;9:315–326.
9. Cardoza JD, Filly RA, Podrasky AE. The dangling choroid plexus: a sonographic observation of value in excluding ventriculomegaly. AJR Am J Roentgenol 1988;151:767–770.
10. Rovas L, Sladkevicius P, Strobel E, Valentin L. Intraobserver and interobserver reproducibility of three-dimensional gray-scale and power Doppler ultrasound examinations of the cervix in pregnant women. Ultrasound Obstet Gynecol 2005;26:132–137.
11. Borrell A, Costa D, Delgado RD, Martinez JM, Borrell C, Fortuny A. Interobserver variability of midtrimester fetal nuchal thickness. Eur J Obstet Gynecol Reprod Biol 1997;72:27–29.
12. Heiserman J, Filly RA, Goldstein RB. Effect of measurement errors on sonographic evaluation of ventriculomegaly. J Ultrasound Med 1991;10:121–124.
13. Ewigman BG, Crane JP, Frigoletto FD, LeFevre ML, Bain RP, McNellis D. Effect of prenatal ultrasound screening on perinatal outcome. RADIUS Study Group. N Engl J Med 1993;329:821–827.
14. Bennett GL, Bromley B, Benacerraf BR. Agenesis of the corpus callosum: prenatal detection usually is not possible before 22 weeks of gestation. Radiology 1996;199:447–450.

15. Glenn OA, Goldstein RB, Li KC, et al. Fetal magnetic resonance imaging in the evaluation of fetuses referred for sonographically suspected abnormalities of the corpus callosum. J Ultrasound Med 2005;24:791–804.

16. Smith M. Error and variation in diagnostic radiology. Springfield, Ill: Thomas, 1967.

17. Renfrew DL, Franken EA Jr, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992;183:145–150.

18. Guo WY, Chang CY, Ho DM, et al. A comparative MR and pathological study on fetal CNS disorders. Childs Nerv Syst 2001;17:512–518.

19. Levine D, Barnes PD, Madsen JR, et al. Fetal CNS anomalies revealed on ultrafast MR imaging. AJR Am J Roentgenol 1999;172:813–818.

20. Sonigo PC, Rypens FF, Carteret M, Delezoide AL, Brunelle FO. MR imaging of fetal cerebral anomalies. Pediatr Radiol 1998;28:212–222.

21. Quinn TM, Hubbard AM, Adzick NS. Prenatal magnetic resonance imaging enhances fetal diagnosis. J Pediatr Surg 1998;33:553–558.

22. Shinmoto H, Kashima K, Yuasa Y, et al. MR imaging of non-CNS fetal abnormalities: a pictorial essay. RadioGraphics 2000;20:1227–1243.

23. Simon EM, Goldstein RB, Coakley FV, et al. Fast MR imaging of fetal CNS anomalies in utero. AJNR Am J Neuroradiol 2000;21:1688–1698.

24. Smith-Bindman R, Hosmer WD, Caponigro M, Cunningham G. The variability in the interpretation of prenatal diagnostic ultrasound. Ultrasound Obstet Gynecol 2001;17:326–332.

25. Birkelo CC, Chamberlain WE, Phelps PS, et al. Tuberculosis case finding: a comparison of the effectiveness of various roentgenographic and photofluorographic methods. JAMA 1947;133:359–366.

26. Borgstede JP, Lewis RS, Bhargavan M, Sunshine JH. RADPEER quality assurance program: a multifacility study of interpretive disagreement rates. J Am Coll Radiol 2004;1:59–65.

27. Szklo M, Nieto FJ. Epidemiology: beyond the basics. Gaithersburg, Md: Aspen, 1999;375–388.

28. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients medical research. Stat Med 2002;21:2109–2129.