

# A Comparison Among 5 Methods for the Clinical Diagnosis of Fetal Alcohol Spectrum Disorders

Claire D. Coles, Amanda R. Gailey, Jennifer G. Mulle, Julie A. Kable, Mary Ellen Lynch, and Kenneth Lyons Jones

**Background:** Despite the prevalence of fetal alcohol spectrum disorders (FASD) and the importance of accurate identification of patients, clinical diagnosis may not be consistent across sites due to the heterogeneous nature of FASD and the characteristics of different diagnostic systems used. Here, we compare 5 systems designed to operationalize criteria recommended for the diagnosis of effects of prenatal alcohol exposure (PAE). We determined the extent of consistency among them as well as factors that may reduce intersystem reliability. Compared are: Emory Clinic, Seattle 4-Digit System (Diagnostic Guidelines for Fetal Alcohol Spectrum Disorders: The 4-Digit Diagnostic Code, Seattle, WA, University Publication Services, 2004), Centers for Disease Control and Prevention (Fetal Alcohol Syndrome: Guidelines for Referral and Diagnosis, Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA, 2004), Canadian Guidelines (CMAJ, 172, 2005, S1), and the Hoyme Modifications (Pediatrics, 115, 2005, 39).

**Methods:** Subjects were 1,581 consecutively registered patients applying for evaluation at a university-based clinic treating alcohol and drug-exposed children. Records of the multidisciplinary evaluation (pediatric, social, psychological) were abstracted. Diagnostic criteria for all 5 systems were applied, and patients were diagnosed according to each of the systems. We compared results using Cohen's Kappa to evaluate the extent of agreement.

**Results:** Percent of individuals diagnosed with FASD ranged from 4.74% (CDC) to 59.58% (Hoyme). Examination using Cohen's Kappa found modest agreement among systems, particularly when individual diagnoses, Fetal Alcohol Syndrome (FAS), partial FAS (pFAS), and Alcohol-Related Neurodevelopmental Disorder (ARND) were used. Examination of diagnostic criteria found almost perfect agreement on growth (weight; height), with limited overlap for physical features (palpebral fissures, hypoplastic philtrum, upper vermillion) and for neurobehavioral outcomes. Child's race and age influenced agreement among systems, with African American and older children more frequently diagnosed.

**Conclusions:** Results suggest problems in convergent validity among these systems, as demonstrated by a lack of reliability in diagnosis. Absence of an external standard makes it impossible to determine whether any system is more accurate, but outcomes do suggest areas for future research that may refine diagnosis.

**Key Words:** Fetal Alcohol Syndrome, Fetal Alcohol Spectrum Disorders, Diagnoses Phenotype.

THE DEVELOPMENTAL DISORDERS associated with prenatal alcohol exposure (PAE) have been described as fetal alcohol spectrum disorders (FASD), with

*From the Department of Psychiatry and Behavioral Sciences (CDC, AG, JAK, MEL), Emory University School of Medicine, Atlanta, Georgia; Department of Pediatrics (CDC, JAK), Emory University School of Medicine, Atlanta, Georgia; Department of Epidemiology (JM), Rollins School of Public Health, Emory University, Atlanta, Georgia; Department of Human Genetics (JM), Emory University School of Medicine, Atlanta, Georgia; and Department of Pediatrics (KLJ), University of California, San Diego, California.*

*Received for publication December 17, 2015; accepted January 29, 2016.*

*Reprint requests: Claire D. Coles, PhD, Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, 12 Executive Park Drive, Atlanta, GA 30329; Tel.: 404-712-9814; Fax: 404-712-9809; E-mail: ccoles@emory.edu*

*Copyright © 2016 by the Research Society on Alcoholism.*

DOI: 10.1111/acer.13032

fetal alcohol syndrome (FAS) the most severe. The potential negative consequences for FASD over the life span make a clear understanding of the nature of these disorders and the implications of a diagnosis important in working with affected individuals and families (Hanlon-Dearman et al., 2015; Ryan et al., 2006). However, since the first descriptions of FAS 40 years ago (Jones et al., 1973; Lemoine et al., 1968), it has been challenging to develop practical and unambiguous diagnostic criteria (Astley, 2014a; Sokol and Clarren, 1989), largely because of the heterogeneous nature of these disorders. Although the reality and the seriousness of the syndrome are not in doubt, there remains no single test or biomarker that can reliably identify FASD. It is universally agreed that there are 4 criteria that indicate the effects of PAE: (i) evidence of maternal alcohol use in pregnancy, (ii) growth retardation, (iii) presence of physical features associated with alcohol exposure, and (iv) neurodevelopmental deficits. Several groups and agencies have proposed diagnostic systems to operationalize these criteria and improve

identification of alcohol-affected individuals (Astley and Clarren, 2000; Chudley et al., 2005; Hoyme et al., 2005; Stratton et al., 1996). In this paper, we compared 5 such systems, selected because each offered practical methods for the application of the Institute of Medicine's (IOM) criteria (Stratton et al., 1996) or for improvement over it. The majority are well known and used, in some form, by multiple clinicians in North America. These systems were also chosen because the diagnostic algorithm or criteria are not only published but are well specified. Therefore, these systems could be operationalized and applied to an existing database of consecutively registered patients applying for evaluation at a university-based clinic specializing in treatment of alcohol- and drug-exposed children. The systems compared were: (i) Emory-Fetal Alcohol Center Clinical Criteria, Atlanta, Georgia (Blackston et al., 2005; Coles et al., 1997), included as this is the system used in the original data collection. This system is an adaptation of the IOM (Stratton et al., 1996) recommendations and, as information has not been published previously about its use, forms and information are included in Data S1 and S2. The major differences from other systems are that it uses a weighted checklist of 40 items to obtain a "dysmorphology score" to rate the physical effects of alcohol, yielding a score from "0" to "57" rather than sentinel facial features alone. Scores of 10 and 20 were used in the current study to evaluate the effect of different thresholds on diagnostic outcomes. These thresholds were chosen as "10" is 2 standard deviations from the nonexposed mean in our exposure cohorts (e.g., Coles et al., 1997) and "20" is based on clinical judgement. Neurodevelopment is based on results of cognitive assessment using standardized measures and does not include measurement of emotional/behavioral or adaptive function. The authors will provide more detailed information on request, (ii) Seattle-based 4-Digit System (Astley, 2004, 2006, 2013; Astley and Clarren, 2000), (iii) Centers for Disease Control and Prevention (CDC) FAS Guidelines for Referral and Diagnosis, July 2004 (Bertrand et al., 2004), (iv) Canadian Guidelines (Chudley et al., 2005; Looock et al., 2005), and (v) Hoyme Modification of IOM Criteria (Hoyme et al., 2005). Table 1 compares the characteristics of these systems. A review of this table will indicate the similarities and some of the differences in the application of the 4 diagnostic criteria in the identification of FAS, partial FAS (pFAS), and Alcohol-Related Neurodevelopmental Disorder (ARND), which is diagnosed when alcohol exposure is confirmed and there is evidence of neurodevelopmental deficit without physical features. Two of these systems, the CDC and the 4-Digit Code, explicitly do not include ARND as a category; the 4-Digit Code does propose methods for indicating neurobehavioral deficit in the presence of alcohol exposure (Astley, 2006). Neurobehavioral Disorder-Prenatal Alcohol Exposure (ND-PAE), another diagnosis recently proposed for further evaluation in the Diagnostic and Statistical Manual, 5th Edition (DSM-5) (Kable et al., 2015), was not considered in this analysis as it was not extant when these systems were

proposed and criteria for its application have not been integrated into these systems.

## MATERIALS AND METHODS

The study was carried out using consecutively registered patients, ages 0 to 21, who applied for a multidisciplinary diagnostic evaluation from 1995 through 2011 at a university-based clinic specializing in the care of children with prenatal alcohol and drug exposure. The protocol for inclusion in the study was approved by Emory University School of Medicine's Internal Review Board. Patients whose guardians agreed to abstraction of their medical records at the initial appointment were included in the database. Children all received a multidisciplinary assessment of physical, psychological/developmental, and demographic/social status. Information was obtained from children's birth records as well as social service and educational records. Presence of dysmorphic features was evaluated by a pediatric geneticist. This examination using a checklist to quantify dysmorphic features yielded a weighted score that quantified the effects of PAE. (Please see Data S1 and S2 for the forms used to evaluate children along with an explanation of diagnostic methods for the Emory System.) Threshold scores of 10 and 20 on this measure were evaluated in this study. Both alcohol-related and other diagnoses were recorded.

The original sample comprised 1,884 children with a median age of 5.22 years. After exclusions, 1,581 cases remained in the analysis. Most cases were excluded because of missing data elements necessary for analyses, with 117 cases (6.9%) excluded because of a known genetic condition and consent withdrawn in 3 cases (0.17%).

To carry out a comparison of the 5 systems, it was necessary to create algorithms using operational definitions of each system's own diagnostic criteria. In some cases, this was easy to do (i.e., birthweight  $\leq 10$ th percentile). In other cases, it was more difficult, particularly for the neurobehavioral criteria (e.g., "clinical judgement"). Criteria used in this process were defined based on published material. When the published material was not specific enough, authors responsible for a given system were requested to provide further information (Hoyme, 2014; S. Astley, personal communication; J. L. Cook, personal communication; W. O. Kalberg, personal communication).

The 4 criteria used by these 5 systems to make diagnoses of FAS, pFAS, and ARND were operationalized. In the case of the 4-Digit Code, diagnostic outcomes of neurodevelopmental deficits with known alcohol exposure were included. In all cases when norms were required (e.g., palpebral fissure length [PFL]), we used those recommended by the diagnostic systems themselves. Criteria included: alcohol exposure (confirmed: Yes; No); growth retardation (as defined by each system and using the World Health Organization growth norms recommended by the CDC, 2010; WHO, 2006); physical features (defined as per each system including the recommended norms); and neurobehavioral deficit (also defined as per each system). There were substantial differences among systems in how physical features were identified and in the definition of neurobehavioral deficits.

For the Emory, Canadian, Hoyme, and CDC systems, algorithms were created to identify presence or absence of each of the 4 criteria for FASD diagnosis: (i) alcohol use – Yes = 1; No = 0, (ii) growth retardation – Yes = 1; No = 0, (iii) physical features – Yes = 1; No = 0, and (iv) neurobehavioral deficit – Yes = 1; No = 0. For each system, these data were assembled into 4-digit numbers, where each digit is a placeholder for the presence or absence of a feature. (For example, a code of "0000" would indicate the absence of all 4 features, while "1111" would indicate the presence of all 4 features.) Each code was binned into its respective diagnostic category according to the guidelines for the system being evaluated. Individuals having all 4 criteria or 3 other criteria in the

**Table 1.** Criteria for Clinical Classification of FASD (FAS, pFAS, and ARND) by 5 Systems

Diagnostic system	Prenatal exposure	Dysmorphic features	Growth retardation	Neurodevelopment	Partial FAS	"ARND"/FASD	Comments
1. EMORY FAS Clinic, 2000 <sup>a</sup>	1. Documents mother's alcohol consumption patterns during pregnancy. 2. Does not accept "hearsay" evidence.	1. Uses weighted scoring system to characterize dysmorphic features. Uses "10" and "20" as Cuto Score for diagnosis versus "sentinel features." 2. Features based on Jones and Smith, 1973, and results from longitudinal exposure sample.	1. Weight or length/height $\leq 10$ th percentile at birth or diagnosis. 2. Must rule out other causes for growth retardation.	1. Head circumference $\leq 10$ th percentile under 3 years or medical signs. 2. Cognitive deficits based only on developmental and cognitive testing. 3. Behavioral reports/checklists not accepted.	1. Requires that exposure be confirmed. 2. Two of the 3 other criteria. Does not necessarily have to have the 3 facial features.	1. Requires that ETOH exposure be confirmed. 2. Cognitive deficits w/o physical features.	1. Extensive sample both longitudinal and clinical. 2. Data on both reliability and validity, but data largely unpublished.
2. 4-Digit System, Astley and Clarren, 2000 <sup>b</sup>	Rankings based on risk of exposure from: Level 1, no risk with confirmed lack of exposure through level 4, confirmed high levels of alcohol exposure.	Relies on 3 "sentinel" features identified through statistical analysis of clinical sample. Features must be objectively measured: PBL 2 SD below M; Lip and philtrum = 4/5 on Lipometer.	Rankings qualifying effects range from weight and height/length $\leq 10$ th percentile (level 2) through at $\leq 2$ nd percentile (level 4).	1. For "4" level must have head circumference $< 3$ rd percentile- "microcephaly." Also abnormal brain structure, seizure disorder or IQ $\leq 70$ . 2. For other ranks, 3 + domains of behavior at $> 2$ SD below mean. Includes behavioral syndromes and descriptors.	Based on results of 4-Digit analysis, produces 9 categories, from no effects through FAS. Includes Atypical FAS and FAS Phenocopy.	Does not use this term, but includes categories that are analogous w/o attributing effects to ETOH (e.g., "Neurobehavioral Disorder, alcohol exposed").	1. Large number of possible classifications. 2. Lipometer and computerized measurement operationalize features. 3. Widely adopted as it reduces the "subjectivity" in diagnosis of many features.
3. Centers for Disease Control and Prevention (CDC) and Fetal Alcohol Syndrome Task Force, 2004 <sup>c</sup>	1. Confirmed prenatal Alcohol Exposure. 2. Unknown prenatal alcohol exposure.	Documentation of 3 facial features, shortened palpebral fissures, flattened philtrum, thin upper lip.	Weight or length $\leq 10$ th percentile at any one point in time.	1. Head circumference $\leq 10$ th percentile. 2. Neurological signs. 3. Functional deficits in developmental and cognitive testing, either globally ( $< 2$ SD) or in 3 areas of development ( $< 1$ SD).	This report notes that there is no adequate scientific evidence for the creation of other diagnostic categories at this time.	Not included.	1. Restricts to 3 dysmorphic features. 2. Cognitive criteria are very broad. 3. Emphasizes importance of differential diagnosis with other diseases causing dysmorphic features.

Continued

Table 1. (Continued)

Diagnostic system	Prenatal exposure	Dysmorphic features	Growth retardation	Neurodevelopment	Partial FAS	"ARND"/FASD	Comments
4. Canadian System, 2005 <sup>d</sup>	Documentation of mother's alcohol consumption patterns during index pregnancy.	Relies on 3 "sentinel" features recommend by 4-Digit system. Palpebral fissure length (PFL) measured differently.	Weight or length/height $\leq 10$ th percentile.	Head circumference at $\leq 3$ rd percentile. Multiple cognitive and behavioral indicators including "clinical judgement."	1. Requires that exposure be confirmed. 2. Requires the 3 facial features. 3. Requires evidence of neurocognitive deficits/behavior. 4. Does not use term "pFAS" unless there is evidence of neurodevelopmental deficit.	Confirmed ETOH plus impairment in 3 CNS domains.	1. Attempts to reconcile IOM and 4-Digit Code systems. 2. Accepts the 3-feature dysmorphism criteria. 3. Very broad criteria for neurobehavioral evidence. 4. Restrictive for infants and preschoolers.
5. Hoyme Modifications, 2005 <sup>e</sup>	1. Confirmed "excessive" alcohol use by mother. 2. Unconfirmed exposure.	1. Uses weighted scoring system to characterize extent of dysmorphia. Score not used in diagnosis. 2. 2 of 3 "sentinel" features required.	Weight or length/height $\leq 10$ th percentile.	For FAS, accepts only medical conditions indicating CNS damage and head circumference $\leq 10$ th percentile included as "microcephaly"; for FASD, accepts very broad and vague descriptions of behavior deficits.	1. Exposure may or may not be confirmed. 2. Requires 2 facial features. 3. Requires evidence of neurocognitive deficits/behavior or growth.	Confirmed ETOH and one of: Structural CNS deficits; HC $< 10$ th percentile; or pattern of cognitive impairments.	1. Growth criteria maximize "sensitivity" while reducing "specificity." 2. Uses clinical judgement of "pattern" associated with FASD for neurobehavioral criteria.

<sup>a</sup>Blackston et al., 2005; Coles et al., 1997.<sup>b</sup>Astley, 2004, 2013; Astley and Clarren, 2000.<sup>c</sup>Bertrand, et al., 2004.<sup>d</sup>Chudley et al., 2005; Look et al., 2005.<sup>e</sup>Hoyme et al., 2005.



absence of alcohol confirmation were diagnosed as "FAS." Partial FAS was defined as having 2 other criteria in the presence of alcohol confirmation (except for the Hoyme system that allows a diagnosis of pFAS without alcohol exposure, given the presence of certain other criteria). Similar criteria were employed consistent with each system's requirements for the diagnosis of ARND. For ARND, confirmation of alcohol was required when specific neurodevelopmental/behavioral characteristics were present but physical characteristics were not. For the Seattle 4-Digit code system, presence of symptoms was not rated by presence or absence, but instead was rated on a 4-digit scale, fully consistent with instructions for clinical coding for this system (Astley, 2004). These data were then assembled into 4-digit codes that correspond to the codes for the 4-digit system. These 4-digit codes were then translated into diagnostic categories as recommended by this system (Astley, 2004).

Subsequently, the percentages diagnosed in each FASD category for each system were calculated. Chi-Square and Fisher's exact tests were computed to compare outcomes. Cohen's Kappa (Cohen, 1960) was used to compare the degree to which diagnostic systems were consistent in their ratings. To interpret Cohen's Kappa results, we followed the Landis and Koch (Landis and Koch, 1977) convention, in which Kappas of 0 to 0.20 indicate "slight," 0.21 to 0.40 indicate "fair," 0.41 to 0.60, "moderate," 0.61 to 0.80, "substantial," and 0.81 to 1.00 "almost perfect" agreement.

## RESULTS

### *Sample Characteristics*

Table 2 shows the demographic characteristics of the clinic sample used for this study. For the 1,581 cases retained in the sample, the mean age at diagnosis was 5.98 years (SD = 3.99).

Table 3 shows the percentage of 1,581 cases diagnosed for the categories, namely FAS, pFAS, "ARND," and No Diagnosis, by each of 5 diagnostic systems, namely Emory-20, the Seattle 4-Digit Code, Canadian (Canada), Hoyme Modifications of IOM criteria (Hoyme), and the Centers for Disease Control and Prevention (CDC). Note that the CDC system only classifies FAS and does not include other categories of FASD and the 4-Digit system does not use the term ARND or impute causation by alcohol in those cases. Table 3 shows percentages only. In understanding these results, it is important to recognize that cases classified in one system may be placed in a different category in another system. These discrepancies are shown graphically in Fig. 1.

### *Degree of Agreement Among Systems*

**FAS Diagnosis Only.** As the CDC system only includes FAS as a diagnostic category, we initially analyzed agreement among the 5 systems using only the FAS category (Table 4a). In these analyses, the CDC system shows fair to moderate agreement with the Canadian (0.506), Hoyme (0.407), and Emory-20 (0.329) systems and only slight agreement with the 4-Digit Code (0.097). The Emory-20 and the Hoyme systems show moderate agreement (0.535), while the Canadian system generally shows only "slight" agreement with all the systems except the CDC. Similarly, the 4-Digit Code system shows only slight agreement with all other systems.

**Alcohol-Affected (FASD) Versus No Diagnosis.** If all of 3 possible diagnoses are collapsed (FAS+pFAS+ARND), 4 of the 5 systems can be compared (Comparison with the CDC system is not appropriate; the 4-Digit explicitly does not use the term ARND, but states that their categories of

**Table 2.** Demographic, Growth, Caregiver, and Cognitive Characteristics

Variable	N	Range	Mean $\pm$ SD
Age (years)	1,706	0.08–29.00	6.06 $\pm$ 4.03
Gestational Age (weeks)	1,612	23–43	37.59 $\pm$ 3.27
Full Scale IQ	1,178	30–141	84.49 $\pm$ 16.50
Infant Development Quotient	383	50–145	84.31 $\pm$ 16.56
Birth weight (%)	1,573	0–100	34.35 $\pm$ 27.22
Birth length (%)	1,232	0–100	39.59 $\pm$ 29.79
Birth HC (%)	1,112	0–100	36.13 $\pm$ 31.20
Current weight (%)	1,695	0–99.9	47.36 $\pm$ 32.76
Current height (%)	1,696	0.1–99.9	39.75 $\pm$ 31.20
Current HC (%)	1,692	0.1–99.9	51.98 $\pm$ 33.44

Variable	N	%
Gender		
Male	1,006	59.0
Female	700	41.0
Race/Ethnicity		
White	713	41.8
African American	788	46.2
Asian	3	0.2
Native American	4	0.2
Hispanic	26	1.7
Other (Including Biracial)	154	9.0
Reason for Referral <sup>a</sup>		
FAS or R/O FAS	1,642	96.2
Dysmorphic Features	53	3.1
R/O effect of other substance	602	35.3
Developmental Delay/LP	667	39.1
Significant Behavior Problem	784	46.0
Sexual Acting Out/Sexual Abuse	77	4.5
Adoption/Foster Care Planning	17	1.0
Legal	4	0.2
Other	113	6.6
Primary Caregiver		
Biological Mother	108	6.3
Biological Father	34	2.0
Grandparent	397	23.3
Other Relative	159	9.3
Caseworker	3	0.2
Foster Parent	555	32.5
Adoptive Parent	420	24.6
Other	27	1.6
FASD diagnosis R/O due to other Medical Condition	117	6.9

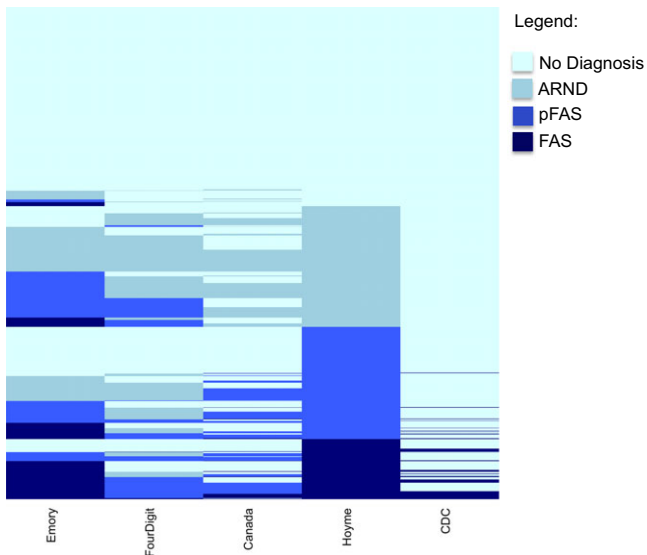
<sup>a</sup>Each case may have multiple reasons for referral.

R/O = Rule Out; LP = Learning Problem.

**Table 3.** Percent of Alcohol-Related Diagnoses by Diagnostic System (n = 1,581)

System	FAS	pFAS	ARND	Any alcohol Dx	No diagnosis
Emory-20	13.73%	16.13%	15.94%	45.79%	54.21%
4-Digit Code	0.25%	12.97%	24.29%	37.51%	62.49%
Canada	1.83%	10.31%	13.03%	25.17%	74.83%
Hoyme	12.21%	22.83%	24.54%	59.58%	40.42%
CDC	4.74%	N/A	N/A	4.74%	95.26%

Figures in the fourth column are a sum of the first 3 columns.



**Fig. 1.** Map of diagnostic outcomes for each system. Each case ( $n = 1,581$ ) is represented as a single line with the color in each column corresponding to the diagnosis with dark blue representing FAS; royal blue, pFAS; grey, ARND; and light blue, no diagnosis. Note that an individual case may be diagnosed with FAS in one system, pFAS in another, and even no diagnosis in a third.

static encephalopathy/alcohol exposed and neurobehavioral disorder/alcohol exposure are equivalent; Astley, 2014b). Table 4b shows the Kappa scores indicating the degree to which there is agreement in diagnosing an individual as alcohol affected between each of the systems. Agreement between the 4-Digit Code, the Hoyme Modifications, and the Emory-20 system is in “moderate” range, while the Canadian system is in the “fair” to “moderate” range. The greatest agreement is between the Emory and 4-Digit code (“substantial agreement”), while the least is between the Canadian and the Hoyme systems (“fair”) (Table 4b).

**Agreement Across FASD Spectrum.** When the diagnostic outcomes are considered individually, as FAS, pFAS, “ARND,” or No Diagnosis, agreement is substantially lower, generally in the “fair” to “moderate” range, with Kappas ranging from 0.31 to 0.46 (Table 4c).

These results suggest that there is a moderate to substantial degree of agreement among systems when individuals are classified as alcohol affected or not but that when diagnosis is more specific, the concordance is much lower. To evaluate the contribution of specific criteria to the diagnostic outcomes, we also examined the concordance of the growth, physical features, and neurobehavioral criteria among systems. In addition, we evaluated outcomes by whether or not alcohol use was confirmed to determine how knowledge of alcohol exposure drove diagnosis.

**Specific Diagnostic Criteria.** In examining specific criteria, growth was found concordant at an “almost perfect” level, suggesting that differences among systems could not be attributed to this diagnostic feature (Table 5a). In contrast,

**Table 4.** Cohen’s Kappa Statistics for Diagnostic Agreement Between Systems. a. FAS Diagnosis Only. b. Any Alcohol-Related Diagnosis (FAS, pFAS, and ARND). c. Individual FASD Diagnosis (FAS, pFAS, or ARND)

a						
System	4-Digit code	Canada	Hoyme	Emory-20	Emory-10	CDC
Emory-20 <sup>a</sup>	0.022	0.160	0.535		0.517	0.329
Emory-10 <sup>a</sup>	0.007	0.063	0.369	0.517		0.170
4-Digit Code		0.117	0.036	0.022	0.007	0.097
Canada	0.117		0.172	0.160	0.063	0.506
Hoyme	0.036	0.172		0.535	0.369	0.407
CDC	0.097	0.506	0.407	0.329	0.170	

b					
System	4-Digit code	Canada	Hoyme	Emory-20	Emory-10
Emory-20 <sup>a</sup>	0.69	0.48	0.61		0.81
Emory-10 <sup>a</sup>	0.55	0.36	0.56	0.81	
4-Digit Code		0.63	0.57	0.69	0.55
Canada	0.63		0.35	0.48	0.36
Hoyme	0.57	0.35		0.61	0.56

c					
System	4-Digit code	Canada	Hoyme	Emory-20	Emory-10
Emory-20 <sup>a</sup>	0.46	0.31	0.40		0.58
Emory-10 <sup>a</sup>	0.25	0.24	0.39	0.58	
4-Digit Code		0.46	0.35	0.46	0.35
Canada	0.46		0.37	0.31	0.37
Hoyme	0.35	0.37		0.40	0.39

<sup>a</sup>Emory-10 refers to a total score of “10” on a physical symptoms checklist, while Emory-20 refers to a total score of 20. Both systems were included in the analysis to estimate the effect of making the physical features criterion more or less stringent.

agreement of physical (facial) features ranges from 0.029 (essentially “no agreement”) between the 4-Digit Code and both the Hoyme and Emory systems to 0.80 (“almost perfect”) between the Canadian and the CDC systems (Table 5b). Discrepancy with the Emory system, which relies on a broader spectrum of physical features than the other systems, was expected; however, discordance among the other systems was not expected, as they all rely on occurrence of the same 3 facial features (palpebral fissure length [PFL], hypoplastic philtrum, thinned upper vermillion). Further analyses suggested that system differences in norms for PFL and the absence of reliable norms for children under 6 years of age contributed to discrepancies. Finally, concordance among systems for neurobehavior showed significant discrepancies (Table 5c). This criterion is defined differently among systems and agreement ranges from 0.15 (“slight”) between the Canadian and the Hoyme systems and the CDC and Hoyme systems to a high of 0.53 between the Emory and CDC systems and 0.54 between the Canadian and 4-Digit systems (both in the “moderate” range).

The probability of diagnosis as a function of confirmed alcohol exposure was also calculated. We used the variable “Alcohol use confirmed: Yes or No.” If the answer was “Yes,” it was assumed that the individual had been exposed. If “No,” it was assumed that alcohol use was unknown,

although nonuse could not be assumed, given the type of referrals that the clinic received. In all systems, knowledge of alcohol use greatly increased diagnosis of FASD (Table 6).

**Demographic Factors That Might Influence Diagnosis.** To determine the extent to which demographic characteristics of the sample (which has fewer White children and more young children than many other samples) may have influenced results, we carried out analyses of outcomes in relation to sex, race, and age. There were few sex differences in diagnosis, although there were significant differences in likelihood of receiving the diagnoses of pFAS and ARND in older boys

in 2 systems (i.e., Canada,  $\chi^2_{(3)} = 13.69$ ,  $p < 0.003$ ; and Hoyme,  $\chi^2_{(3)} = 9.67$ ,  $p < 0.02$ ). Race differences were noted in 3 systems (Emory,  $\chi^2_{(15)} = 29.73$ ,  $p < 0.01$ ; 4-Digit,  $\chi^2_{(15)} = 38.08$ ,  $p < 0.001$ ; and Hoyme  $\chi^2_{(15)} = 29.61$ ,  $p < 0.01$ ), with African Americans being diagnosed more frequently. Age differences were examined by dividing the sample into 6 categories, namely (i) 0 to 2 years, 11 months ( $n = 420$ ), (ii) 3 years to 5 years, 11 months ( $n = 491$ ), (iii) 6 years to 8 years, 11 months ( $n = 337$ ), (iv) 9 years to 11 years, 11 months ( $n = 175$ ), (v) 12 years to 14 years, 11 months ( $n = 108$ ), and (vi) 15 years and older ( $n = 45$ ). In general, we observed that infants and young children were diagnosed less frequently by all systems in which we had power to detect this trend (for the CDC system, the total N of those diagnosed was too small to measure effectively). These differences were most extreme using the Canadian system ( $\chi^2_{(15)} = 144.08$ ,  $p < 0.000$ ). To summarize these differences, we evaluated odds ratios for diagnosis of children older than 6 versus those younger for any FASD, and these are shown in Table 7.

## DISCUSSION

All 5 of these systems were designed to characterize the developmental effects of PAE. Our assumption in beginning this analysis was that application of these systems to the diagnosis of the same group of children at risk for FASD would lead to generally similar results. Doing so would demonstrate convergent validity in the diagnosis of disorders associated with PAE. We anticipated that there might be some discrepancies related to specific criteria. For example, we believed that those systems that require 3 facial features (i.e., 4-Digit, Canadian, CDC) rather than 2 (i.e., Hoyme) for diagnosis of FAS would identify fewer cases. We also anticipated that the Emory System, which relies on a broader range of features, might identify more children. However, these 5 systems are only moderately similar in diagnostic outcomes. This is true for the absolute percentages of cases diagnosed (Table 3) and also, at an individual level: the same individual may receive a different diagnosis depending on the method used (Fig. 1). It might be assumed that this is usually a matter of degree. That is, while one system would render a diagnosis of pFAS, another would diagnosis FAS. However, there are cases in which one system would call an individual

**Table 5.** Cohen's Kappa Statistics for Agreement Among Systems for Specific Diagnostic Criteria. a. Growth Variables Only. b. Facial (Physical) Features Only. c. Neurobehavior Only

a System	4-Digit code	Canada	Hoyme	Emory-20/10	CDC
Emory	0.91	0.85	1.00		1.00
4-Digit Code		0.75	0.91	0.91	0.91
Canada	0.75		0.85	0.85	0.81
Hoyme	0.91	0.85		1.00	1.00
CDC	0.91	0.85	1.00	1.00	

b System	4-Digit code	Canada	Hoyme	Emory-20	Emory-10	CDC
Emory-20 <sup>a</sup>	0.028	0.23	0.39		0.23	0.28
Emory-10 <sup>a</sup>	0.008	0.053	0.43	0.23		0.073
4-Digit Code		0.30	0.29	0.28	0.008	0.21
Canada	0.30		0.15	0.23	0.05	0.80
Hoyme	0.029	0.15		0.39	0.43	0.22
CDC	0.21	0.80	0.22	0.28	0.073	

c System	4-Digit code	Canada	Hoyme	Emory-20/10	CDC
Emory	0.23	0.21	0.18		0.53
4-Digit Code		0.54	0.45	0.23	0.21
Canada	0.54		0.15	0.21	0.22
Hoyme	0.45	0.15		0.18	0.15
CDC	0.21	0.22	0.15	0.53	

<sup>a</sup>Emory-10 refers to a total score of "10" on a physical symptoms checklist, while Emory-20 refers to a total score of 20. Both systems were included in the analysis to estimate the effect of making the physical features criterion more or less stringent.

**Table 6.** Percent of Sample Receiving FASD Diagnoses when Alcohol Use is Confirmed/Unconfirmed ( $N = 1,581$ )

System	Alcohol confirmed ( $n = 822$ )	Alcohol unconfirmed ( $n = 759$ )	Odds ratio	95% confidence interval	p-value
Emory	81.5%	20.7%	48.96	35.7–67.1	<0.000
4-Digit	71.8%	0.4%	640.86	204.2–2011.8	<0.000
Canadian	47.8%	0.7%	138.15	56.7–336.4	<0.000
Hoyme	87%	29.9%	15.66	12.3–20.25	<0.000
CDC <sup>a</sup>	6.8%	2.5%	2.85	1.68–4.84	<0.000

<sup>a</sup>FAS Diagnosis only ( $n = 75$ ).

**Table 7.** The Odds of Receiving a FASD Diagnosis for Children Over Versus Under 6 Years of Age in Each System

System	Odds ratio	95% Confidence interval	p-value
Emory	0.955	0.78–1.17	NS
4-Digit	1.289	1.05–1.58	<0.02
Canada	2.786	2.21–3.52	<0.000
Hoyme	1.696	1.38–2.09	<0.000
CDC <sup>a</sup>	1.347	0.85–2.14	NS

<sup>a</sup>Diagnosis confined to FAS; N is limited ( $n = 75$ ).

FAS, while another would not give any diagnosis. Concordance is somewhat better when the different diagnostic categories are collapsed into FASD versus No Diagnosis, although agreement is still in the moderate range for the most part. These findings raise, but do not answer, questions about the usefulness of the current diagnostic discriminations.

Substantial agreement among systems would have inspired confidence that all systems accurately identify a similar, underlying condition, or set of relationships, and that the methods recommended were appropriate for such identification. Given these results, however, we cannot be comfortable making these assumptions without further evaluation. In addition, in the absence of an external standard, despite the obvious discrepancy among systems, we cannot say that one system is better or worse in identifying the effects of prenatal alcohol exposure.

When viewed in retrospect, this discrepancy is inevitable. These 5 methods have each been developed by individuals from different disciplines with different ideas regarding the diagnosis of FASD; some have used a more or less rigorous definition of FASD; and not all have used the same reference data for normal physical measurements.

Consistency in diagnosis is important to move this field forward. Our results indicate that that improvement in diagnostic consistency could be made by further attention to specification of physical features and neurobehavior. Standardization of these criteria, including agreement on the underlying norms to be used, could lead to improved application and greater reliability. Improvement in the identification by physical feature could probably be achieved through empirical examination of the physical features used in the diagnosis as well as those typically evaluated in a physical examination of children suspected of being alcohol affected. Using empirical methods it might be possible to identify characteristics that could refine the current criteria. For instance, different systems use 2 or 3 facial features, or in the case of Emory, results of a weighted checklist. However, the decisions to select these characteristics have not, for the most part, been made using empirical evidence but rather rely on clinical judgement. It would be important to know, for instance, whether both thinned upper vermillion and hypoplastic philtrum contribute independently to the validity of a diagnosis. In addition, there may be other features that could add to the reliability/validity of the diagnosis that are not currently being measured. In the new Canadian Guidelines (Cook et al., 2015), the authors report the utility of other facial features in the identification of FASD.

The problems associated with definition of neurobehavioral outcomes in FASD are more comprehensive than can be discussed in this summary paper and are related to the problems observed in all classification of psychiatric and behavioral disorders (Kraemer, 2014; Regier et al., 2013). However, it is evident that these 5 systems employ a wide variety of measures and thresholds in meeting the neurobehavioral criterion and there is inconsistency among

them. As it is this area of development that is most likely to affect the child's life course as well as the most likely to be influenced by demographic and social factors, it is important that further attention should be directed at understanding the contribution of PAE to developmental and behavioral outcomes and some consistency in standards be obtained.

Other questions raised by these data concern the utility of the various subcategories that are used in the diagnosis. Is it really possible to discriminate FAS from pFAS and ARND using the current methods? Should the term FASD be substituted, given that current diagnostic methods are more convergent when all the diagnoses are combined? Interestingly, the current revision of the Canadian Guidelines adopts this approach (see Cook et al., 2015).

Finally, it is evident from these results that children's age and race influence diagnostic decisions. While some of these differences may result from factors like socioeconomic status and philosophical views on the accuracy of infant diagnosis, it is important that these issues be examined further to allow more effective diagnosis in clinical settings.

### *Limitations*

The most significant problem for the interpretation of these results is that there is no "gold standard" against which to evaluate them, and without such an external standard, we cannot say that one system is better or worse in identifying the effects of PAE. There is no biomarker currently and expert opinion has been the standard up until this time. A second type of external indicator would be knowledge of prenatal alcohol exposure, particularly if a dose/response measure could be used. Unfortunately, most individuals present for diagnosis without accurate information on maternal alcohol use, and valid and reliable dose information is particularly difficult to obtain.

Another potential limitation is that the sample used was patients applying consecutively for services and, therefore, self-selected. Any biases in this process will be preserved in the data. The characteristics of the individuals in the sample reflect the characteristics of the region in which they live rather than the general population of the United States. Thus, as Atlanta is a predominantly African American city, the sample is 47% African American, while there are fewer Hispanic children relative to the general population. Another possible bias is socioeconomic status (SES). The majority of the children referred were Medicaid recipients and many were in foster care, a pattern consistent with the strong relationship between maternal substance abuse and alcoholism and the likelihood of the child entering foster care or being adopted. The child's development certainly may be negatively impacted by such factors and such impact may influence the results of developmental and behavioral tests. Another important characteristic of the sample is that the majority of children were not in the care of their biological parents. While this is commonly the case in clinical samples



of FASD, other samples concerned with the diagnosis of FASD (exposure samples, surveillance carried out in first grade class rooms) are selected from different populations and, therefore, may have different demographic profiles.

While not a bias, the age of the children included in this study should be considered in interpreting results. With a median age of 5.98 years, the majority of children are infants and preschool children. Some of the systems studied (i.e., 4-Digit; Canada) intentionally limit the diagnosis of young children because they believe that it is not possible to accurately characterize their neurodevelopmental characteristics. However, this choice may result in underidentification of children during the age range when intervention can be most effective.

## CONCLUSION

Kraemer in her discussion of clinical diagnostic reliability (Kraemer, 2014) discriminates between disorder, which is the condition within the patient, and diagnosis, which is the “informed opinion” of the professional. There is unquestionably a developmental disorder associated with PAE (Riley et al., 2011), but like many neurobehaviorally based disorders, it can be difficult to diagnosis and this is particularly true at the milder end of the range of effects. Although the physical characteristics associated with PAE (growth retardation and dysmorphic features) are signs that may signal the disorder, FASD is a developmental psychopathology emerging from the impact of alcohol on the developing brain and the subsequent interaction with the environment that produces characteristic behavior that sometimes may be difficult to discriminate from other factors.

The 5 systems evaluated here have all contributed to the understanding of this process, and their comparison highlights some of the next steps that will be required to refine and standardize methods for identification and diagnosis in the future. By identifying areas that are discrepant, we can see better where more research and standardization are required. For instance, there are some obvious steps that can be taken to refine the classification of physical features using existing data. Age differences in diagnostic consistency also emerge as a focus for future research. The results of such research will allow for more efficient identification and treatment of alcohol-affected individuals in the future. The current study points to the importance of developing a consensus diagnosis for this common disorder based on empirical evidence and suggests directions for future research that will improve identification of children with FASD.

## ACKNOWLEDGMENTS

NIH/NIAA-Administrative Supplement to U01 AA019879 and DHR 427-93-05050762 Fetal Alcohol and Drug Screening Project. The authors declare that they have no conflict of interest.

## REFERENCES

- Astley S (2004) Diagnostic Guidelines for Fetal Alcohol Spectrum Disorders: The 4-Digit Diagnostic Code, 3rd edn. University Publication Services, Seattle, WA.
- Astley SJ (2006) Comparison of the 4-digit diagnostic code and the Hoyme diagnostic guidelines for fetal alcohol spectrum disorders. *Pediatrics* 118:1532–1545.
- Astley SJ (2013) Validation of the fetal alcohol spectrum disorder (FASD) 4-Digit Diagnostic Code. *J Popul Ther Clin Pharmacol* 20:e416–e467.
- Astley SJ (2014a) Invited commentary on Australian fetal alcohol spectrum disorder diagnostic guidelines. *BMC Pediatr* 14:85.
- Astley SJ (2014b) The value of a FASD diagnosis (2013). *J Popul Ther Clin Pharmacol* 21:81–105.
- Astley SJ, Clarren SK (2000) Diagnosing the full spectrum of fetal alcohol-exposed individuals: introducing the 4-digit diagnostic code. *Alcohol Alcohol* 35:400–410.
- Bertrand J, Floyd RL, Weber MK, O'Connor M, Riley EP, Johnson KA, FAS/FAE NTFO (2004) Fetal Alcohol Syndrome: Guidelines for Referral and Diagnosis. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA.
- Blackston RD, Coles CD, Kable JA (2005) Evidence for Severity of Dysmorphology in Fetal Alcohol Syndrome and Direct Correlation With Developmental, Behavioral, Social and Educational Outcomes and to Psychotropic Medications. David Smith Dysmorphology Meeting, Iowa City, Iowa.
- Centers For Disease Control And Prevention (2010) *WHO Child Growth Standards are Recommended for Use in the U.S. for Infants and Children 0 to 2 Years of Age*. Growth Charts, National Center for Health Statistics [Online]. Available at: [http://www.cdc.gov/growthcharts/who\\_charts.htm](http://www.cdc.gov/growthcharts/who_charts.htm). Accessed October 28, 2014.
- Chudley AE, Conry J, Cook JL, Looock C, Rosales T, Leblanc N, Public Health Agency Of Canada's National Advisory Committee On Fetal Alcohol Spectrum Disorder (2005) Fetal alcohol spectrum disorder: Canadian guidelines for diagnosis. *CMAJ* 172:S1–S21.
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Measur* 20:37–46.
- Coles CD, Platzman KA, Raskind-Hood CL, Brown RT, Falek A, Smith IE (1997) A comparison of children affected by prenatal alcohol exposure and attention deficit, hyperactivity disorder. *Alcohol Clin Exp Res* 21: 150–161.
- Cook JL, Green CR, Lilley MM, Anderson SM, Baldwin ME, Chudley AE, Conry JL, Leblanc N, Looock CA, Lutke J, Mallon BF, McFarlane AA, Temple VK, Rosales T (2015) Fetal Alcohol Spectrum Disorder (FASD): Guidelines for diagnosis across the lifespan. *CMAJ* 188:191–197.
- Hanlon-Dearman A, Green CR, Andrew G, Leblanc N, Cook JL (2015) Anticipatory guidance for children and adolescents with Fetal Alcohol Spectrum Disorder (FASD): practice points for primary health care providers. *J Popul Ther Clin Pharmacol* 22:e27–e56.
- Hoyme HE (2014) Personal communication with author, December 11, 2014, RE: diagnosis of FAS.
- Hoyme HE, May PA, Kalberg WO, Kodituwakku P, Gossage JP, Trujillo PM, Buckley DG, Miller JH, Aragon AS, Khaole N, Viljoen DL, Jones KL, Robinson LK (2005) A practical clinical approach to diagnosis of fetal alcohol spectrum disorders: clarification of the 1996 institute of medicine criteria. *Pediatrics* 115:39–47.
- Jones KL, Smith DW (1973) Recognition of the fetal alcohol syndrome in early infancy. *Lancet* 302:99–100.
- Jones KL, Smith DW, Ulleland CN, Streissguth P (1973) Pattern of malformation in offspring of chronic alcoholic mothers. *Lancet* 1:1267–1271.
- Kable JA, O'Connor MJ, Olson HC, Paley B, Mattson SN, Anderson SM, Riley EP (2015) Neurobehavioral disorder associated with prenatal alcohol exposure (ND-PAE): proposed DSM-5 diagnosis. *Child Psychiatry Hum Dev* 47:335–346.
- Kraemer HC (2014) The reliability of clinical diagnosis: state of the art. *Annu Rev Clin Psychol* 10:11–130.

- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174.
- Lemoine P, Harousseau H, Borteyru JP (1968) Les enfants de parents alcooliques: anomalies observees, a propos de 127 cas. *Ouest Med* 21:476–482.
- Loock C, Conry J, Cook JL, Chudley AE, Rosales T (2005) Identifying fetal alcohol spectrum disorder in primary care. *CMAJ* 172:628–630.
- Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramotor SJ, Kuhl EA, Kupfer DJ (2013) DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry* 170:59–70.
- Riley EP, Infante MA, Warren KR (2011) Fetal alcohol spectrum disorders: an overview. *Neuropsychol Rev* 21:73–80.
- Ryan DM, Bonnett DM, Gass CB (2006) Sobering thoughts: town hall meetings on fetal alcohol spectrum disorders. *Am J Public Health* 96:2098–2101.
- Sokol RJ, Clarren SK (1989) Guidelines for the use of terminology describing the impact of prenatal alcohol exposure on the offspring. *Alcohol Clin Exp Res* 13:597–598.
- Stratton K, Howe C, Battaglia F (1996) *Fetal Alcohol Syndrome: Diagnosis, Epidemiology, Prevention and Treatment*. National Academy Press, Washington, DC.
- World Health Organization (2006) *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development* [Online]. Available at: [http://www.who.int/childgrowth/standards/technical\\_report/en/](http://www.who.int/childgrowth/standards/technical_report/en/).

## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

**Data S1.** FASD dysmorphology checklist.

**Data S2.** Emory neurodevelopmental exposure clinic FAS diagnostic criteria.