Title

Comparison of three systems for the diagnosis of fetal alcohol spectrum disorders in a community sample.

Authors

Coles, Claire

Bandoli, Gretchen

Kable, Julie

et al.

Publication Date

2023-02-01

DOI

10.1111/acer.14999

Peer reviewed

Published in final edited form as:

Alcohol Clin Exp Res (Hoboken). 2023 February ; 47(2): 370–381. doi:10.1111/acer.14999.

Comparison of three systems for the diagnosis of Fetal Alcohol Spectrum Disorders in a community sample

Claire D. Coles[1], **Gretchen Bandoli**[2], **Julie A. Kable**[1], **Miguel del Campo**[2], **Michael Suttie**[3,4], **Christina D. Chambers**[2]

[1]Departments of Psychiatry and Behavioral Sciences and Pediatrics, Emory University School of Medicine, Atlanta, GA

[2]Departments of Pediatrics and Family Medicine and Public Health, University of California San Diego School of Medicine, La Jolla, CA

[3]Department of Women's & Reproductive Health, University of Oxford, Oxford England

[4]Big Data Institute, University of Oxford, Oxford England

Abstract

Background: It is estimated that 1 to 5% of children in the United States may be affected by prenatal alcohol exposure while only a small percentage are so identified in clinical practice. One explanation for this discrepancy may be the way in which diagnostic criteria are operationalized.

Methods: To evaluate the extent to which three commonly used systems for the diagnosis of Fetal Alcohol Spectrum Disorder (FASD) consistently identified children in a community sample, data from the Collaboration on Fetal Alcohol Spectrum Disorders Prevalence (COFASP) study were re-analyzed. In the dataset, there were 2325 children with variables necessary to allow diagnosis by three systems commonly used in North America. These systems were 1) that used by COFASP, which is a revised modification of the Institute of Medicine's recommendations, 2) the 4-Digit Code, and 3) the most recent Canadian Guidelines.

Results: Among these three systems, 408 children were classified as FASD, 208 by the CoFASP system, 319 by the 4-Digit Code, and 28 by the Canadian Guidelines. To determine the degree of association among these classifications, the Fleiss Multirater Kappa measure of agreement was applied finding that agreement varied from slight to fair, among systems.

Conclusions: These results indicate a lack of consistency in these approaches to diagnosis. Discrepancies result from differences in specifying the criteria used to define the diagnosis, including growth, physical features, neurobehavior and alcohol-use thresholds. The question of their relative accuracy cannot be resolved without reference to a measure of validity that does not currently exist and this suggests the need for a more empirically based diagnostic schema.

Corresponding author: Claire D. Coles, PhD, Department of Psychiatry and Behavioral Sciences, 12 Executive Park Dr, NE, Emory University School of Medicine, Atlanta, GA 30329, Phone: 404 712 9814, ccoles@emory.edu.

The authors have no conflicts of interest to report.

Keywords

Fetal Alcohol Syndrome; Fetal Alcohol Spectrum Disorders; Diagnostic System; Prenatal Alcohol Exposure

The negative outcomes of prenatal alcohol exposure (PAE) were brought to the attention of professionals and public in the 1970's through a series of articles (e.g., Jones, et al, 1973; Lemoine et al, 1968) describing the fetal alcohol syndrome (FAS). Since that time, the extent of impact of PAE on development has been explored and currently is thought of as comprising a range of fetal alcohol spectrum disorders (FASD) with FAS being only the most severe and easily identified. There are a number of approaches to the description of FAS as well as the other conditions that make up FASD both in North America and worldwide. Those following the nomenclature suggested by the Institute of Medicine (Stratton et al, 1996) also identify a partial fetal alcohol syndrome (pFAS) in which not all of the physical characteristics are present. When effects are manifested only through function and behavior, the condition is described as Alcohol Related Neurodevelopmental disorder (ARND) (Coles, et al, 2020). Recently, using this system, the first epidemiological survey of a community sample done in the United States (May, et al., 2018) suggested that, even among first grade students who were not being referred for clinical service, alcohol effects ranging along this spectrum could be identified in 1 to 5 %. However, this is not the only system for categorizing the outcomes of prenatal alcohol exposure on the child either in North America or in other parts of the world. The purpose of this paper is to compare this system to two others in wide use to determine the degree to which there is consistency among them (see Table 1).

Despite what this recent study suggested about prevalence, FASD is rarely diagnosed in general clinical practice in the United States (Chasnoff, Wells, & King, 2015) denying services to many who should be receiving care to address immediate developmental concerns, assure more positive outcomes and to avoid potentially negative long-term consequences (Coles, et al, 2022; Kable, Mehta, & Coles, 2021). Similarly, it has been difficult to identify the elements necessary to institute a national surveillance system for FASD. There are many reasons for the under-identification of FASD both clinically and through public health surveillance systems, but certainly among them are challenges posed by diagnostic methods.

Initially (Jones and Smith, 1973) and in most subsequent conceptualizations of FASD, there has been agreement on the general characteristics of the syndrome. As noted by the Institute of Medicine in 1996 (Stratten, et al, 1996), these include: 1) Growth retardation[1]; 2) Dysmorphic features, particularly facial features, 3) Impact on the Central Nervous System (CNS) as manifested through neurophysiological or neurodevelopmental deficits; and 4) Evidence of Alcohol Exposure during Pregnancy. Despite this general agreement, there are many ways in which these criteria are interpreted that can affect who is diagnosed.

[1]But note that the current Canadian Criteria (Cook, et al., 2016) and the Australian Criteria (Bower & Elliot, 2016), have eliminated the Growth requirement. Australian scientists and clinicians are currently revising their system (N. Reid, personal communication) and it is for that reason that it has not been included in this paper.

In a previous study, (Coles, et al, 2016), five systems for the diagnosis of FASD were compared, including that used at the Emory University Clinic, a modification of the IOM system, the 4-Digit Diagnostic Code (Astley, et al., 2004), the Centers for Disease Control and Prevention recommendations (Bertrand, et al., 2004), the Hoyme, et al. modifications (2005), and the initial Canadian Guidelines (Chudley, et al.,2005). In a sample of 1,706 clinically referred children ranging in age from birth to 21, the study found that agreement among systems was fair to moderate at best, whether considering absolute number of cases diagnosed or at the individual level as the same person might receive a different diagnosis depending on the system used. When the individual criteria were examined, growth was most likely to be consistent among systems, while physical features, in particular palpebral fissure length (PFL), and neurobehavior were least consistent. These inconsistences resulted from differences in both thresholds and norms for the characterization of physical features as well as the way in which neurodevelopmental deficits were conceptualized.

Since that time, two of these systems have been revised, with Hoyme, et al. (2016) publishing a revision of those criteria based on standards developed for the recent epidemiological study (May, et al, 2019). Also in 2016, Cook, et al. published new recommendations in the form of a new set of Canadian Guidelines. (See Table 1 for a description of the criteria for those systems considered in this analysis.)

In 2019, Hemingway, et al. compared these two new systems with the existing Seattle 4-Digit Code and the Australian 2016 system. This comparison also employed a clinical sample of children (N=1,392) who had received services from the diagnostic clinic at the University of Washington. As found previously, the proportion of the patient group diagnosed with FASD varied significantly among these different diagnostic systems (ranging from 79% for the 4-Digit Code to 16% in the new Canadian system.) The authors attributed the differences to the ways in which the diagnostic criteria were defined (e.g., definition of alcohol use; specification of facial features and use of particular norms; inclusion/exclusion of growth, not including children less than 6 years as diagnosable; and including or excluding moderate neurodevelopmental dysfunction as diagnostic of FASD.)

Although there are many systems for FASD diagnosis both in North America and worldwide, in the current analysis, we compared the specific criteria used in the epidemiological study carried out by the Collaboration on Fetal Alcohol Spectrum Disorders Prevalence research consortium (CoFASP) (i.e., Hoyme, et al. 2016) to those of the 4-Digit Code (Astley, et al., 2004) and the new Canadian Guidelines (Cook, et al., 2016). The current study is unique in employing a Community rather than a clinically-referred sample of children. As such, it is expected that far fewer children will qualify for FASD than in previous comparison studies. The sample is also restricted to first grade children thus focusing the age range in a way that previous studies have not. We hypothesized that this examination would further highlight the way in which application and specification of diagnostic criteria affect the number and kind of children identified with FASD and would allow exploration of the utility of these systems for epidemiological research and public health surveillance.

Methods

This is a secondary analysis of data initially collected by CoFASP. (See May, Chambers, et al, 2018 for a description of the methodology used in that study.) In 2010, the National Institute on Alcohol Abuse and Alcoholism (NIAAA) initiated the CoFASP research consortium. The consortium used active case-ascertainment in collecting data in a community sample in four locations between 2010 and 2016 to estimate the prevalence of fetal alcohol spectrum disorders (FASD) among first grade children in the United States. From the sample of 3,397 available from this epidemiological study, a subsample of 2325 was selected whose data included all items necessary for the diagnosis of FASD via the three diagnostic systems being compared. Data were obtained directly from investigators who include the authors of this paper. Data in this paper are limited to individuals who will be included in the public-use dataset that will be released in the near future.

Study Population.

Participants were first grade children identified through participating schools in 4 communities in the United States. Identification was done in several stages with the initial stage identifying children at risk due to physical status (< 25th percentile) and/or teacher/parental concerns about development as well as low risk “controls”. Parents and caregivers and children’s teachers also participated by completing questionnaires. Characteristics of the study population are shown in Table 2.

Diagnosis of FASD.

For the purposes of the epidemiological study, CoFASP specified a set of criteria to operationalize the identification of FAS, pFAS, and ARND based both on previous research and clinical experience. Subsequently, a more generalized version of these criteria was proposed by Hoyme et al (2016) for use in the diagnosis of FASD in clinical settings.

As shown in Table 1, assessments of children were based on the four criteria typically used for the diagnosis of FASD: growth (current percentile for height, weight, and head circumference); physical features (via a 47-item dysmorphology examination completed by a pediatric dysmorphologist); neurodevelopment, ascertained using a battery of standardized tests (administered by school psychologists or study psychometrists), and PAE. PAE during the time around conception (pre-recognition) and at three timepoints during pregnancy was assessed through maternal or collateral questionnaires administered by trained interviewers. For more information on study design and enrollment, please see the original publication (May, et al., 2018). The study was approved by the University of California San Diego Human Research Protections Program.

Measurement of physical characteristics of PAE.

For the purposes of the current analysis, these included growth and facial features, measured by a pediatric dysmorphologist. Growth included current height, weight, and head circumference which were measured by study staff. Percentiles for growth (height, weight, head circumference) relied on norms recommended by the Centers for Disease Control and Prevention (CDC.gov. CDC Growth Charts [accessed 2019 September 1]). The CoFASP

study used the 10th percentile or less to meet criterion, while the 4-Digit Codes uses the 3rd percentile or less to define "severe" effects and the 10th or less for milder effects. The Canadian system does not use growth as a criterion.

All of the systems compared in this paper use the same three cardinal facial features (i.e., sentinel facial features [SFF]), specifically palpebral fissure length (PFL), philtrum and vermillion. The philtrum and vermillion were ranked from 1 to 5 based on a Lipometer with those ranked 4 or 5 considered to be consistent with the effects of prenatal exposure. In the original data collection for CoFASP the Lipometers used by the Hoyme, et al (2016) methodology were employed. As noted in Astley, et al. (2017), there are some differences in the Hoyme guides from those employed by the 4-Digit Code and this difference in measurement method may have the potential to alter the rankings assigned to the child. Another potential for difference in rankings results from the use of photographs for measurement in the 4-Digit Code system, while the CoFASP study used expert judgement by the dysmorphologists following physical examination of the child. Since this was a secondary data analysis, we could not apply Astley, et al.'s (2004) methods for evaluation of the philtrum and upper lip in this context.

PFL was calculated from the size of this feature as measured for each child by the dysmorphologist and percentiles were based on the norms recommended by each diagnostic system. The CoFASP used the Hall system (Hall, Froster-Iskenius, & Allanson, 1989). Therefore, the PFL percentiles had to be recalculated for both other systems. The Canadian System uses the Canadian PFL norms (Clarren, et al., 1998), and the 4-Digit Code system currently recommends the use of the Scandinavian norms (Stromland, Chen, Norberg, Wennerstrom & Michael, 2010); however, it also requires that African-American and mixed race children be ranked based on the norms recommended by Iosub, et al., (1985); thus, the 274 children who were identified as African-American or mixed race were separately calculated. Based on the recommendations of these individual systems, PFL was recoded and ranked according to each system. Children with percentiles consistent with criteria in Table 1 or lower were considered "positive" for this feature (that is, meeting the cutoff for alcohol effects).

Effects on the Central Nervous System.

The third area assessed was the potential impact on the CNS. This impact was measured either in terms of neuroanatomy and electrophysiology (i.e., microcephaly, diagnosis of seizure disorder, other physical signs) or as functional outcomes assumed to be associated with effects on nervous system (i.e., cognition, behavior, diagnosis of mental health disorders). Medical information, including diagnoses of physical and mental disorders, was obtained when the child was examined by the dysmorphologist. Cognitive and behavioral status were obtained through psychoeducational evaluation of the child and from questionnaires completed by the parent/caregiver. (see Table 3 for a description of these tests.) This information was then used to meet the specific criteria required for each diagnostic system (Table 1). Since standardized tests were used in assessing cognition and behavior, it was possible to specify scores at 1SD, 1 ½ SD and 2 SD from the mean and to use these results to meet the criteria specified by each system.

Alcohol Exposure.

Whether or not the child had been exposed to alcohol prenatally was determined based on the responses to a comprehensive questionnaire by the birth mother or a collateral source. On this questionnaire, the respondent provided information about quantity and frequency of use during the pre-pregnancy recognition period and each trimester. Information was also obtained about binge patterns of use during pregnancy. Information was also available about legal problems associated with alcohol use and any substance use treatment that occurred during the pregnancy. This information was used to meet the alcohol use criteria recommended by each system.

In the CoFASP study women were classified as 1) Not responding; 2) Not drinking during pregnancy; 3) Drinking during pregnancy below the criterion threshold (Any Alcohol) and 4) Meeting criteria for drinking at a level placing the child at risk (Alcohol Criterion). To fall into category “4”, one or more of the following conditions had to be met based on information obtained from the biological mother or a reliable collateral source (e.g., family member).

- **a.** 6 or more standard drinks per week for 2 or more weeks during pregnancy
- **b.** 3 or more standard drinks per occasion on 2 or more occasions during pregnancy
- **c.** Documentation of alcohol-related social or legal problems in proximity to (prior to or during) the index pregnancy (e.g., history of multiple citations for driving while intoxicated or history of treatment for an alcohol-related condition) (Hoyme et al, 2016).

The 4-Digit Code system uses the following categories to rank alcohol use: 1) No risk, in which alcohol use is confirmed to be absent; 2) Unknown risk, in which alcohol use is unknown; 3) Some risk, in which alcohol use is confirmed but at a level less than high risk or not known; and 4) High risk, in which alcohol use is confirmed and the exposure pattern is consistent with medical literature putting the fetus at “high risk”. Since a specific standard was not specified, for the purpose of this study, those meeting the CoFASP criteria for High Risk alcohol use were considered to have met the 4-Digit standard. Therefore, children were classified for this analysis as follows: 1) No risk, if their mother or collateral indicated that there was no alcohol use during pregnancy; 2) Unknown risk, if the mother/collateral did not respond to the questionnaire; 3) Some risk, if the respondent indicated that there was alcohol use during the pregnancy but it did not meet the Alcohol Criterion as specified by CoFASP; and 4) High Risk, if the respondent qualified for the CoFASP Alcohol Criterion.

For application to the Canadian Guidelines (Cook, et al., 2016), the quantity/frequency data on alcohol use during each period of pregnancy were reanalyzed and the sample was reclassified consistent with those criteria. These categories were: 1) No risk (no or limited alcohol use during pregnancy); 2) Unknown (no information about use/mother did not respond to questionnaire); and 3) High risk, meeting the following conditions:

- **a.** 7 drinks per week or more either pre-recognition or in pregnancy.
- **b.** Binge: 4+ drinks=binge; Two binges during pregnancy or pre-recognition are necessary.

c. Other social indicators, including alcohol treatment during pregnancy and legal/social or medical problems related to drinking during the pregnancy.

Data Analysis

After diagnostic classifications were made, to make comparison among these three systems, reduced categories (No Diagnosis, Unknown, ARND, pFAS and FAS) were created for the CoFASP and 4-Digit System and further reduced, for comparison with the Canadian Guidelines, by combining the pFAS and FAS categories since both these categories require facial features. pFAS often does not meet the growth criteria and, as a result, appears consistent with the Canadian classification. It is recognized that, due to differences in the way in which these systems classify cases, that these categories are not completely analogous. Fleiss' Kappa (Fleiss, 1971) was employed to examine the consistence in categorization between systems. Based on the recommendations of Lands and Koch (1977) and Bakeman and Quera (1997), the following rules were applied to determine the degree of relationship: 0, No relationship; 0.1 to .2, Slight agreement, .21 to .4, Fair agreement, .41 to .60, Moderate agreement, .61 to .80, Substantial agreement, .81 to 1.0, Perfect agreement.

Results

Demographic Measures and Outcomes.

In the CoFASP data set, 2325 had data available that allowed diagnostic criteria associated with each of the three systems to be applied. Table 2 (above) shows the demographic characteristics of these individuals and their families. Not all participants responded to all questions so the "N" associated with each characteristic is provided.

Diagnosis based on each system.

The CoFASP study's original classification of children resulted in 14 separate categories describing the method of diagnosis as well as the outcome (i.e., with or without confirmed alcohol exposure; with cognitive or behavioral deficit). To allow the comparison among systems in this analysis, these were collapsed into 4 categories: FAS, pFAS, ARND and no diagnosis.

In the 4-Digit system (Astley, 2004), Growth, Facial features, CNS and Alcohol are each ranked from least impaired (None, Unlikely, No Risk) to most at risk (Severe, Definite, High Risk) on a 4 point scale based on the criteria shown in Table 1. The resulting rankings are then combined into one of a possible 256 numeric scores ranging from 1111 to 4444. These scores are then consolidated into one of 22 unique Clinical Diagnostic Categories represented by letters (A through V). Finally, for the purpose of this analysis, these letters were grouped into diagnoses representing Fetal Alcohol Syndrome (FAS), partial Fetal Alcohol Syndrome (pFAS) and Alcohol Related Neurodevelopmental Disorder (ARND), as well as a number of other descriptive categories that are not represented by these diagnoses.

When the 4-Digit system (Astley, 2004) was applied to the 2325 cases included in this study, participants were initially classified into the anticipated 22 categories with 2004 falling into the NonFASD categories. Two individuals were classified as FAS (i.e., FAS,

Alcohol Exposed; FAS, Alcohol Exposure Unknown), 63 as partial FAS (i.e., Partial FAS, Alcohol Exposed; Sentinel physical findings/Static Encephalopathy, Alcohol Exposed; Sentinel physical findings/Neurobehavior Disorder, Alcohol Exposed) and 254 met criteria consistent with ARND (i.e., Static Encephalopathy, Alcohol Exposed and Neurobehavioral Problems, Alcohol Exposed). Thus, a total of 319 individuals were classified as alcohol affected using this system.

The Canadian Guidelines (Cook, et al, 2016) yield only 4 possible diagnostics outcomes, No diagnosis, At Risk, FASD without significant facial features (FASD w/o SFF), and FASD with SFF (FASD w SFF). In the Canadian system, the initial criterion applied is the Alcohol Exposure which can be “no”, “yes” or “unknown”. If this is ‘no”, then the person cannot be diagnosed with FASD. If, “unknown”, then the presence of SFF is evaluated and if this is “yes”, then the CNS criterion is determined. If both (SFF and CNS) are present, then the individual can be classified as FASD with SFF. If fewer than three SFF are present then there is no diagnosis in the absence of alcohol. If Alcohol exposure at sufficient levels is confirmed, then the SFF criterion is applied in the same manner. With the presence of SFF and CNS, the diagnosis is FASD with SFF. If the three facial features are not present, and the CNS criterion is met, the diagnosis is FASD without SFF. Without the CNS criterion being met, there is no diagnosis of FASD possible. Finally, there is an “at risk” category for those who are younger than 6 years or who have not had neurodevelopmental testing. As there were no such children in this sample, this category was not included in the analysis.

Based on the description of these categories and the method for calculating them, the FASD w/o SFF can be considered analogous to the ARND category used by CoFASP and Hoyme et al (2016) and the following category used by the 4-Digit system, Static Encephalopathy, Alcohol Exposed. The FASD with SFF can be thought of as consistent with FAS and partial FAS in CoFASP classifications and the FAS, Alcohol Exposed; FAS, Alcohol Exposure Unknown and the Partial FAS, Alcohol Exposed; and Sentinel physical findings/Static Encephalopathy, Alcohol Exposed in the 4-Digit Code system.

Diagnosis of FASD.

In the original study, using the CoFASP criteria, 208 (8.9%) of children in this subsample of 2325 were identified as having one of the FASD diagnoses (see Table 4), while using the Seattle 4-Digit code criteria, 319 (13.7%) individuals were so identified. Finally, using the Canadian criteria, 28 (1.2%) were diagnosed. In total, 408 individuals (17.5%) were diagnosed by one of these systems. Table 4 shows the numbers and percentages in each of the categories for each system and compares the numbers between each system pair. In the CoFASP Study, 24 were diagnosed as FAS, 98 as pFAS, 86 as ARND and 1953 as not alcohol affected; 165 could not be classified. In contrast, using the 4-Digit code system, 2 children meet criteria for FAS, 63 for those categories that are analogous to pFAS, and 254 for the categories that are analogous to ARND. Finally, 3 children were identified using the Canadian system as having FASD with significant facial features and 25 as having FASD without such features.

For the purpose of comparing the degree of association among these different systems using this data set, the different diagnostic classifications were collapsed into 4 categories for the

CoFASP and 4-Digit Systems. These were: Not FASD, ARND, pFAS and FAS for both the CoFASP and 4-Digit Code while the Canadian system allowed only 3 categories, Not FASD, FASD without SFF and FASD with SFF. The Fleiss Multirater Kappa measure of agreement (Fleiss, 1971) was used to determine the extent to which systems agreed in their diagnostic assessment. Table 5 presents the results of these analyses, which were done among all three groups and individually between each set of groups.

As can be noted, with reference to the guidelines for the strength of relationships (Bakeman & Quera, 2011; Landis & Koch, 1977), agreement among all three of these systems can be described as slight to fair. Agreement between the Canadian system and either of the other two systems is in the no relationship to slight range, while the agreement of the CoFASP and 4-Digit methods is in the fair range overall although the agreement on the diagnosis of FAS, itself, is in the slight range.

As a final analysis, the three systems were compared based on whether they classified participants in any of the FASD categories or not. This provided a comparison of their similarity for FASD “yes” or “no”. As can be observed, the Kappa of .236 is in the “fair” range. However, comparing individual pairs indicates a disagreement between the Canadian system and both others that results in reduced agreement. The Kappa for the COFASP and 4-Digit Code for “any FASD” is .400, in the “fair” range. In contrast, the agreement of the Canadian Guidelines with the CoFASP is .122 (“slight”) and with the 4-Digit is .057 (“slight”).

Discussion

Consistent and reliable diagnosis of conditions is essential both for the care of affected individuals and for public health policy and planning. An accurate diagnosis of fetal alcohol spectrum disorders would allow more effective treatment of patients presenting for care and would support medical and educational policy planning for prevention and intervention. However, this remains only a goal as there are multiple and sometimes inconsistent methods employed around the world. In the current study, to examine the implications of these differences, the criteria employed by three different systems designed to identify and diagnose FASD were applied And as was found in the previous studies that employed clinical samples (Coles, et al., 2016; Hemingway, et al. 2019), there was only slight to fair agreement among these systems when they were applied to the same children.

There are significant differences in the number of individuals being identified as alcohol-affected among these systems with the 4-Digit Code the most inclusive, with 13.7% of the sample falling into one of the categories consistent with FASD. This percentage is 8.9% for the CoFASP system and only 1.2% for the Canadian Guidelines. In addition, there are discrepancies in the proportion of cases meeting criteria for FAS, pFAS and ARND. Nor are the same individuals being identified consistently in the same categories across systems. That is to say, a person might be classified as FAS by one system and as No Diagnosis in another.

These observations of discrepancies in classification are borne out by the analysis of agreement (Kappas, Table 5) that demonstrates a high degree of unreliability among these methods. At best, agreement is only “fair” and, at worst, it approaches zero. It would be beneficial both to clinicians and scientists to know which of these systems, if any, is most valid in its characterization of FASD. However, as no “gold standard” exists against which any of these systems can be measured, it is impossible to determine whether one system is more effective than another in its ability to identify FASD. The lack of agreement, itself, raises questions about the assumptions used in characterizing the disorder that should be a focus of research in the future.

Why do these systems not measure in the same way?

There are at least two reasons for the inconsistencies in these outcomes. First, and probably the most obvious, are discrepancies in the specification of the criteria that are used to define FASD, that is to say, the algorithms. For instance, growth reduction is required in two of the systems and not in the third. Another difference is that the CoFASP system required two of the so-called significant facial features (SFF) for a “positive” rating, while the other two systems required all three of these cardinal features. A third difference is in the levels used as cutoffs for growth and other features. Some systems require the 10th percentile, some the 3rd, while the 4-Digit essentially allows both for some features leading to different types of categorization. Finally, the way in which the diagnostic algorithms are implemented may exclude a child in one system and not in another. For instance, by using the PAE criterion as the initial screen, the Canadian system excludes children who are being included in the CoFASP system which places more weight on physical features and may assume exposure based on the presence of other factors.

At another level are alternative methods for defining elements making up these criteria. For instance, differences in the norms used as the basis for identification, particularly for PFL, can lead to discrepancies in meeting criterion for this feature. For example, when using the Iosub “norms” (1985), as recommended by the 4-Digit Code, none of the African-American and mixed-race children (N= 274) assessed were above the mean for the norming sample (mean Z score=−1.07, SD=0.22, range −1.71 to −.43), a highly unlikely event that suggests that these norms may not be representative since this group typically has a larger PFL size (Coles, et al., 2016). Taking this same group of individuals and examining the mean percentile for the sample compared for each set of norms we find the following: Iosub norms: 36.1st percentile; Hall (CoFASP norms): 18.4th percentile; Canadian norms: 54.1st percentile and Scandinavian (4-Digit) norms: 42.7th percentile. Thus, the Hall norms are likely to yield the highest number of individuals with “small” PFL and the Canadian norms the lowest. This finding suggests that improvement in diagnostic consistency may be achieved by identifying (or creating) an accurate set of norms that is representative of the population and applying them in a standard fashion.

The methods for meeting the neurobehavioral criterion vary among systems, with obvious differences in qualification for diagnosis. Some (i.e., the Canadian Guidelines) require that standardized test scores be ≤2 standard deviations below the mean in three domains of function. Since the COFASP requires deficits of ≤1 ½ standard deviations, this results

in a difference that will clearly limit the number of cases qualifying for diagnosis in the Canadian versus the COFASP system. The 4-Digit Code is the most comprehensive, identifying those at ≤ 2 standard deviations as "severely" affected but those between 1 and 2 standard deviations as "mildly" affected and including both within the "range" of those with FASD.

Similarly, the difference in how risky drinking is defined will change the classification of individuals. This is demonstrated when we compare the numbers meeting the risky alcohol use level for the CoFASP (10.3%) and Canadian (6.3%) systems. The change from 6 to 7 drinks a week and the higher level for "binge" drinking resulted in fewer women meeting this criterion. Ideally, there would be well-defined thresholds for alcohol use known to be associated with specific outcomes. However, such thresholds have not been identified, and it is possible that they cannot be since it is likely that there are individual differences in susceptibility to the effects of exposure. Nevertheless, a consensus among those creating diagnostic systems on this criterion would improve reliability among systems.

Study limitations.

Interpretation of the results of this study must take into account its limitations. This was a secondary analysis of existing data such that only the CoFASP (or Hoyme, et al. 2016) criteria were administered prospectively. Ideally, the other systems would have been carried out prospectively as well as is typically done in making a diagnosis and such an approach would be an appropriate goal for future research. Another possible limitation is that the sample section, as carried out by CoFASP may have influenced results as a proportion of children were recruited as "high risk" based on their physical status (≤25th percentile for growth) or concerns about development. However, a random sample of children was also recruited and it is not clear that "enriching" the sample in this way should affect the consistency of diagnosis across systems. Finally, as this was a cohort study, growth restriction was based solely on current growth measures and did not include information about birthweight and length as would be ideal in making diagnostic decisions.

How can diagnostic systems be improved?

To date, the development of diagnostic methods has relied on "expert opinion", and this practice has led to a significant difference in how the same individual is categorized. As a result, it could be argued that some of the "cut offs" used for classification are arbitrary rather than anchored in empirical knowledge. The results of this study, like those carried out previously, suggest that it may be time to move beyond such methods and begin to incorporate more psychometrically based approaches. For instance, while there is no doubt that PAE has teratogenic effects, it would be important to tie the criterion for alcohol exposure to data derived from exposure studies that have been able to link specific drinking patterns, including dosage and timing, to specific outcomes both in development and growth. Once this is done, measurement standards could be applied that address the variability that currently leads to such wide discrepancies in outcome. The use of longitudinal, exposure studies to document these effects would allow more accurate measurement of these relationships and allow exploration of the potentially confounding environmental factors

that add to error variance. The development and consistent use of appropriate norms for measuring these features is also essential to improving diagnostic accuracy.

However, regardless of the accuracy of a diagnostic system's normative data and stated inclusion criteria, the results will inaccurate be if the physical assessment is imprecise. All three systems described in this study are reliant on a necessarily subjective evaluation of physical dysmorphia and are typically restricted to including only cardinal features. In the updated guidelines by Hoyme et al. (2016), a dysmorphology scoring system is detailed, providing an analysis of a comprehensive list of minor physical anomalies in children with FAS. Among these, growth metrics include height, weight, and head circumference, with the proportion of the sample below $\leq 10^{th}$ percentile for each measure being in excess of 88%. This list also describes minor physical anomalies which have a high prevalence in their FAS sample such as railroad track ears and clinodactyly, in addition to subtle facial dysmorphia such as midfacial hypoplasia, epicanthal folds and a flat nasal bridge. Likewise, the CoFASP system details a comprehensive checklist of clinically observable physical features which largely overlaps with the Hoyme dysmorphology scoring system. While these physical features are recorded as part of physical assessment, their use for diagnosis and screening is limited. The lack of inclusion in diagnostic systems of these physical features is perhaps due to difficulty of recognition, ultimately requiring trained dysmorphologists to subjectively identify a comprehensive list subtle features.

While inter-rater reliability for quantitative measurements such as PFL in FASD evaluation is shown to be good, the more subjective elements of the face such as lip vermillion and philtrum have shown less reliability (May et al., 2013). The introduction of more objective clinical tools such as craniofacial assessment utilizing 3D imaging could enhance the clinical accuracy of facial measurements, and provide a quantitive approach to identifying more subtle soft tissue deformations such as midfacial hypoplasia and nasal bridge flatness. This approach could provide more detailed assessment of facial traits, and even provide automated fully objective tools which could be utilized for initial screening, or in situations where trained dysmorphologists are unavailable (Fu et al., 2022, Suttie et al., 2017).

The Importance of a reliable and valid method for diagnosis of FASD is widely acknowledged as this is necessary to allow identification and care of affected individuals and support the scientific study of the effects of this teratogen. The results of the current analysis suggest that we have not yet been able to reach this standard and that better and more reliable methods for diagnosis should be a goal for the future. To continue without improving consistency in our approach to this problem is to contribute to confusion about the reliability and validity of this diagnosis which does a disservice to all those who are concerned about the care of affected individuals.

Acknowledgements

This research was supported by the following awards from the National Institute on Alcohol Abuse and Alcoholism (NIH/NIAAA) to Christina D. Chambers (U01AA019879) and Philip May (U01 AA019894) as well as to Gretchen Bandoli (R01AA027785)

References

Achenbach TM & Rescorla LA (2001). Manual for the ASEBA School-Age Forms & Profiles.

Burlington VT: University of Vermont, Research Center for children, Youth, & Families.

Astley SJ. Diagnostic guide for fetal alcohol spectrum disorders: The 4-digit diagnostic Code, 3rd edition University of Washington publication services, Seattle, WA [accessed 2019 September 1] 2004.

Astley SJ, Bledsoe JM, Davies JK, & Thorne JC (2017) Comparison of the FASD 4-Digit Code and Hoyme et al. 2016 FASD Diagnostic Guidelines. Advances in Pediatric Research 4:13. [PubMed: 33409370]

Bakeman R, & Quera V. (2011). Sequential analysis and observational methods for the behavioral sciences. Sequential Analysis and Observational Methods for the Behavioral Sciences. 10.1017/CBO9781139017343.

Beery KE, Beery NA (2004). The Beery-Buktenica Developmental Test of Visual-Motor Integration – 5th Edition. San Antonio, Texas: Pearson Assessment.

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..

Bracken BA (1998). Bracken Basic Concept Scale – Revised. San Antonio Texas: Harcourt Assessment, Inc.

Bower C and Elliott EJ. 2016, on behalf of the Steering Group. Report to the Australian Government Department of Health: "Australian Guide to the diagnosis of Fetal Alcohol Spectrum Disorder (FASD)".

CDC.gov. CDC Growth Charts [accessed 2019 September 1]. (boys) https://www.cdc.gov/growthcharts/data/set2clinical/cj41c071.pdf (girls) https://www.cdc.gov/growthcharts/data/set2clinical/cj41c072.pdf.

Chasnoff IJ, Wells AM, & King L (2015) Misdiagnosis and missing diagnosis in foster and adopted children with prenatal alcohol exposure. Pediatrics, 135 (2), 264–270. [PubMed: 25583914]

Chudley AE, Conry J, Cook JL, Loock 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. [PubMed: 15738468]

Clarren SK, Chudley AD, Wong L, Friesen J, & Brant R (2010) Normal distribution of palpebral fissure lengths in Canadian school age children. Canadian Journal of Clinical Pharmacology, 17(1) e67–78.

Coles CD, Gailey AR, Mulle JG, Kable JA, Lynch ME, Jones KL. A comparison among 5 methods for the clinical diagnosis of fetal alcohol spectrum disorders. Alcoholism Clinical and Experimental Research 2016; 40(5);e1000–1009.

Coles CD, Grant TM, Kable JA, Stoner SA, Perez A, and the CIFASD (2022) Prenatal Alcohol Exposure and Mental Health at Midlife: A Preliminary Report on Two Longitudinal Cohorts. Alcoholism: Clinical and Experimental Research. 46 (2) 232–242. PMID: 35157325. [PubMed: 35157325]

Coles CD, Kalberg W, Kable JA, Tabachnick B, May PA, Chambers CD and the CoFASP (2020). Characterizing Alcohol-Related Neurodevelopmental Disorder (ARND): Prenatal alcohol exposure and the spectrum of outcomes. Alcoholism: Clinical and Experimental Research, 44 (6), 1245–1260, PMID:. [PubMed: 32173870]

Cook JL, Green CR, Lillye CM, Anderson SM, Baldwin ME, Chudley AE, Conry JL, LeBlanc N, Loock CA, Lutke J, Mallon BF, McFarlane, Aam Temple VK, Rosales T(and the Canada Fetal Alcohol Spectrum Disorder Research Network) (2016) Fetal alcohol spectrum disorder: A guideline for diagnosis across the lifespan. CMAJ 188(3) 191–197. [PubMed: 26668194]

Elliot CD (1979). Differential Ability Scales, 2nd Edition. San Antonio, Texas: Pearson Assessment.

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin, 76 (5), 378–382.

Fu Z, Jiao J, Suttie M. and Noble JA, "Facial Anatomical Landmark Detection Using Regularized Transfer Learning With Application to Fetal Alcohol Syndrome Recognition," in IEEE Journal

of Biomedical and Health Informatics, vol. 26, no. 4, pp. 1591–1601, April 2022, doi: 10.1109/JBHI.2021.3110680..

Hall JG, Froster-Iskenius UG, Allanson JE. Handbook of normal physical measurements. New York, USA: Oxford university press 1989.

Hemingway SJA, Bledsoe JM, Brooks A, Davies JK, Jirikowic T, Olson E, & Thorne JC (2019) Comparison of the 4-Digit Code, Canadian 2015, Australian 2016 and Hoyme 2016 fetal alcohol spectrum disorder diagnostic guidelines. Advances in Pediatric Research 6 (2), pages doi:10.35248/2385-4529.19.6.31.

Hollingshead AB (2011). Four Factor Index of Social Status. Yale Journal of Sociology, 8, 21–51.

Hoyme HE, Kalberg WO, Elliott AJ, Blankenship J, Buckley D, Marais AS, et al. Updated clinical guidelines for diagnosing fetal alcohol spectrum disorders. Pediatrics. 2016;138(2):e20154256. [PubMed: 27464676]

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. [PubMed: 15629980]

Iosub S, Fuchs M, Bingol N, Stone RK, Gromishch DS, Wasserman E. (1985) Palpebral fissure length in Black and Hispanic children: Correlation with head circumference. Pediatrics, 75 318–320. [PubMed: 3969333]

Jones KL, Smith DW (1973) Recognition of the fetal alcohol syndrome in early infancy. Lancet 302:99–100.

Kable JA, Mehta PK, and Coles CD. (2021) Alterations in insulin levels in adults with prenatal alcohol exposure. Alcoholism Clinical and Experimental Research, 45 (3), 500–506. 10.1111/acer.14559. [PMID: 33486796]. [PubMed: 33486796]

Korkman M, Kirk U, & Kemp S. (2007). NEPSY-II. San Antonio, Texas: Pearson Assessment.

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics, 33, 159–174. [PubMed: 843571]

Lemoine P, Haroosseau H, Borteryu JP and Menuet JC (1968) Les Enfants de Parents Alcooliques. Anomalies Observee a Propos de 127 cas. Quest Medicale, 21, 476–482.

May PA, Blankenship J, Marais AS, Gossage JP, Kalberg WO, Barnard R, De Vries M, Robinson LK, Adnams CM, Buckley D, Manning M, Jones KL, Parry C, Hoyme HE, Seedat S. Approaching the prevalence of the full spectrum of fetal alcohol spectrum disorders in a South African population-based study. Alcohol Clin Exp Res. 2013 May;37(5):818–30. doi: 10.1111/acer.12033. Epub 2012 Dec 14. PMID: 23241076; PMCID: PMC3610844. [PubMed: 23241076]

May PA, Chambers CD, Kalberg WO, Zellner J, Feldman H, Buckley D, Kopald D, Hasken JM, Xu R, Honerkamp-Smith G, Taras H, Manning MA, Robinson LK, Adam MP, Abdul-Rahman O, Vaux K, Jewett T, Elliott AJ, Kable JA, Askhoomoff N, Faulk D, Arroyo JA, Hereld D, Riley EP, Charness M, Coles CD, Warren KR, Jones KJ, Hoyme HE, and the CoFASP. (2018) Prevalence of Fetal Alcohol Spectrum Disorders in Four Communities of the United States: Results from the Collaboration on Fetal Alcohol Spectrum Disorders Prevalence (CoFASP). Journal of the American Medical Association 6:319 (5), 474–482.

Stratton K, Howe C, Battaglia F. Fetal alcohol syndrome: Diagnosis epidemiology prevention and treatment Institute of medicine. Washington D C National Academy Press 1996.

Stromland K, Chen Y, Norberg T, Wennerstrom K, Michael G. Reference values of facial features in Scandinavian children measured with a range-camera technique. Scand J Plast Reconstr Hand Surg. 1999; 33:59–65.

Suttie M, Wetherill L, Jacobson SW, Jacobson JL, Hoyme HE, Sowell ER, Coles C, Wozniak JR, Riley EP, Jones KL, Foroud T, Hammond P; CIFASD. Facial Curvature Detects and Explicates Ethnic Differences in Effects of Prenatal Alcohol Exposure. Alcohol Clin Exp Res. 2017 Aug;41(8):1471–1483. doi: 10.1111/acer.13429. Epub 2017 Jul 10. PMID: 28608920; PMCID: PMC5563255. [PubMed: 28608920]

Sparrow AS, Cicchetti DV, Balla DA. (2005). Vineland Adaptive Behavior Scales – Second Edition. Bloomington, Minnesota: Pearson Assessment.

Table 1:

Comparison of Criteria in Diagnostic Systems Used for Categorization of FASD

	Diagnostic System		
	CoFASP	**4-Digit**	**Canadian**
Significant Facial Features (SFF)	2 of 3 Facial Features: PFL<10th % Philtrum/Vermillion Ranked 4 or 5	3 Facial Features: PFL<3rd % Philtrum/Vermillion Ranked 4 or 5	3 Facial Features: PFL<3rd % Philtrum/Vermillion Ranked 4 or 5
Growth	Height and/or Weight <10th percentile	Height/Weight Severe: <3rd % Significant 3rd to <10th %	None
Neurodevelopment (CNS)	Evidence of Neurodevelopmental impairment ≤ 1.5 SD Global Cognitive Test;≤ 1.5 SD other neurodevelopmental test or ≥1.5 SD on Behavioral	4 levels: Severe: Physical evidence of neurological impact Significant: Global Cognitive <2 SD; or 3 or more domains <2 SD on standardized tests. Mild: Scores on standardized measures between 2SD and <1 SD. Absent: None of the above	Impairment in 3 or more developmental domains including: Motor skills, Neuroanatomy/ Neurophysiology, Cognition, Language, Academic Achievement, Memory, Attention, Executive Function, Affect regulation, Adaptive Skills, Socialization. Test scores must be <2SD below mean (or above for measures of behavior). Also includes “significant” discrepancy between cognition and other domains. Can also include Diagnosis of mental health conditions.
Alcohol Exposure	≥6 drinks per week for ≥2 weeks in pregnancy; >3dk/occ on ≥2 occ; Documented alcohol-related social/ legal problem during pregnancy; Documentation of intoxication during pregnancy; Positive biomarker during pregnancy; Positive score on Standardized screener.	High Risk: Confirmed Alcohol use with pattern of use placing fetus at risk. Some Risk: Confirmed Alcohol Use at a lower level. Unknown Risk: Alcohol Use is Unknown No Risk: Confirmed lack of alcohol use	Evidence of Heavy alcohol use, i.e.: 7 standard drinks per week. Binge drinking (4–5 standard drinks/occ) on 2 or more occasions.

Table 2:

Demographic, growth, and caregiver characteristics of study sample.

Variable	N	Range	Mean ± SD
Child's Age (years)	2327	4.58–10.8	6.99 ± 0.49
General Conceptual Ability SS[1]	2055	30–147	100.78 ± 12.89
Current weight (% ile)	2325	1–99	53.58± 29.81
Current height (% ile)	2325	1–99	49.75 ± 31.20
Current HC (% ile)	2325	1–99	49.97 ± 31.30
Current Palpebral Fissure Length (left) (% ile)	2327	1–90	27.24±16.29
Philtrum Code (1 to5)	2326	1–5	3.02 ± 0.72
Vermillion Code (1 to 5)	2326	1–5	2.99 ± 0.75

Variable	N	%
Gender		
Male	1186	51.0
Female	1141	49.0
Race/Ethnicity		
White	1761	75.7
African-American	190	8.2
Native American/Alaskan	46	2
Multiracial	180	7.7
Other /Unknown	150	6.5
Maternal Characteristics		
Variable	N	Range	Mean ± SD
Maternal Age Current	1675	22.25–52.83	35.98±5.99
Maternal Years of Education	757	0–24	14.34±2.97
Social Economic Status (Hollingshead)[2]	885	0–67	38.71±12.67
Parity (# living children born)	1698	1–10	2.65±1.19

Variable	N[3]	%
Marital Status: Married/Partnered	1324/1707	78
Alcohol use in pregnancy	250/1682	14.5
Tobacco use in pregnancy	147/818	16.5
Marijuana use in Pregnancy	48/1679	2.9
Prescription drug misuse in pregnancy	20/1560	1.3
Cocaine use in pregnancy	12/1680	0.7
Opiates/Heroin use in pregnancy	7/1682	0.4

[1] SS=Standard Score, M=100; SD=15

[2] Hollingshead, 2011

[3] Number positive/Number responding

Table 3:

Standardized Tests used to Identify Neurobehavioral Effects and Constructs Measured

Test	Construct (s)
Differential Ability Scales, 2nd Edition (DAS-II)[1]	Global Ability/Verbal Ability/NonVerbal Problem Solving/Spatial Ability
NEPSY-II[2] (Specific Subtests: Inhibition/Speeded Naming/ VisuoMotor Precision)	Executive Functioning/ Inhibition/Impulsivity/Cognitive Flexibility Language Fluency/Processing Speed/Sensorimotor Skills/
Beery-Buktenica Developmental Test of Visual-Motor Integration, 5th Edition[3]	Graphomotor Skills/Eye-hand Coordination/Visual Information Processing
Bracken Basic Concepts Scale[4]	School Readiness/Academic Achievement
Vineland Adaptive Behavior Scales, 2nd Edition[5]	Adaptive Function Communication/Socialization/Motor Skills
Child Behavior Checklist (CBCL)[6]	Behavior Problems and Mental Health
Teacher Report Form (TRF)[6]	Behavior Problems and Mental Health

[1] Elliot (1979)

[2] Korkman, Kirk, & Kemp (2007)

[3] Beery & Beery (2004)

[4] Bracken (1998)

[5] Sparrow, Cicchetti, Balla, (2005)

[6] Achenbach & Rescorla (2001).

Table: 4:

Diagnosis of Fetal Alcohol Spectrum Disorders (FASD) by 3 Diagnostic Systems

CoFASP[1]				
FAS	**Partial FAS**	**ARND**	**Not FAS**	**Not Classifiable**
24	98	86	1950	167
4-Digit Code[2]				
FAS	**Partial FAS**	**ARND**	**Not FAS**	**Not Classifiable**
2	63	254	2001	5
Canadian[3]				
FASDwSSF		**FASDw/oSFF**	**Not FAS**	**Not Classifiable**
3		25	2297	

Comparison of Diagnostic Classification Among Groups:				
1) CoFASP and 4-Digit				
	CoFASP System			
4-Digit System	**FAS**	**Partial FAS**	**ARND**	**Not FASD**
FAS	1	1	0	0
Partial FAS	5	28	4	25
ARND	1	15	68	165
Not FASD	17	54	14	1760

2) CoFASP and Canadian				
	CoFASP System			
Canadian	**FAS**	**Partial FAS**	**ARND**	**Not FASD**
FASDwSFF	1	1	0	0
FASDw/oSFF	3	7	9	6
Not FASD	21	90	77	1943

3) 4-Digit and Canadian				
	4-Digit System			
Canadian	**FAS**	**Partial FAS**	**ARND**	**Not FASD**
FASDwSFF	0	0	0	2
FASDw/oSFF	0	7	15	2
Not FASD	2	56	239	1997

[1] Mays, et al., 2019

[2] Astley, et al., 2004

[3] Cook, et al., 2016

X^2_{15}=802, p<.001 Pearson's Correlation: r=−.025, NS

X^2_{10}=540.68, p<.001; Pearson's Correlation: r=.001, NS

X^2_8=304.21, p<.001; Pearson's Correlation: r=.197, p<.001

Table 5:

Agreement on Specific Diagnosis Among Diagnostic Systems Using Fleiss' Multirater Kappa: Total Sample (N=2325)

A. CoFASP and 4 Digit Code (4 Categories each) (Usable N=2158)

Overall Agreement[1]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.364	.017	21.726	.000	.331	.397

Agreement on Individual Categories

		Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Not FASD	.400	.022	18.568	.000	.358	.442
ARND	.356	.022	16.537	.000	.314	.398
pFAS	.325	.022	15.097	.000	.283	.367
FAS	.071	.022	3.314	.001	.029	.114

B. CoFASP and Canadian System (3 categories each) (Usable N=2158)

Overall Agreement[2]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.090	.017	5.357	.000	.057	.124

Agreement on Individual Categories

		Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
.Not FASD	.122	.022	5.687	.000	.080	.165
ARND/FASDw/oSFF	.140	.022	6.506	.000	.098	.182
pFAS/FAS/FASDwSFF	-.013	.022	-.603	.547	-.055	.029

C. 4-Digit Code and Canadian System (3 categories each) (Usable N=2324)

Overall Agreement[3]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.047	.018	2.615	.009	.012	.082

		Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Not FASD	.057	.021	2.746	.006	.016	.098
ARND/FASDw/oSFF	.051	.021	2.436	.015	.010	.091
pFAS/FAS/FASDwSFF	-.015	.021	-.705	.481	-.055	.026

D. All Three systems compared (3 categories each) (Usable N=2158)

Overall Agreement[4]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.216	.010	21.622	.000	.196	.235

Agreement on Individual Categories

		Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Not FASD	.236	.012	19.026	.000	.212	.261
ARND/FASDw/oSFF	.212	.012	17.035	.000	.187	.236
pFAS/FAS/FASDwSFF	.167	.012	13.462	.000	.143	.192

E. All Three Systems compared for Total FASD Diagnosis (all categories combined). versus No Diagnosis

Overall Agreement[5]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.236	.012	19.026	.000	.212	.261

Agreement on Individual Categories

		Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
.00	.236	.012	19.026	.000	.212	.261
1.00	.236	.012	19.026	.000	.212	.261

Degree of agreement Standards: 0-.20 = Slight; .21 to .40 = Fair; .41 to .60 = Moderate; .61 to .80 = Substantial; 81 to 1.00 = Perfect. (Bakeman & Quera, 2011; Landis & Koch, 1977)

[1]Sample data contains 2158 effective subjects and 2 raters.

[2].Sample data contains 2158 effective subjects and 2 raters.

[3]Sample data contains 2324 effective subjects and 2 raters.

[4]Sample data contains 2158 effective subjects and 3 raters.

[5]Sample data contains 2158 effective subjects and 3 raters.