Implementation of a Shared Data Repository and Common Data Dictionary for Fetal Alcohol Spectrum Disorders Research

Andrew D. Arenson[1], Ludmila Bakhireva[2], Christina D. Chambers[2], Christina Deximo[1], Tatiana Foroud[3], Joseph L. Jacobson[4], Sandra W. Jacobson[4], Kenneth Lyons Jones[2], Sarah N. Mattson[5], Philip A. May[6], Elizabeth Moore[7], Kimberly Ogle[5], Edward P. Riley[5], Luther K. Robinson[8], Jeffrey Rogers[1], Ann P. Streissguth[9], Michel Tavares[1], Joseph Urbanski[3], Yelena Yezerets[1], Radha Surya[1], Craig A. Stewart[10], William K. Barnett[1]

[1] Indiana University, University Information Technology Services, Indianapolis, IN 46202, USA. {aarenson, cdeximo, jlrogers, mtavares, yyezert, rsurya, barnettw}@indiana.edu
[2]University of California, San Diego, Department of Pediatrics, La Jolla, CA 92093, USA. {lbakhireva, klyons, chchambers}@ucsd.edu
[3]Indiana University School of Medicine, Department of Medical and Molecular Genetics, Indianapolis, IN 46202, USA. {tforoud, joaurban}@iupui.edu
[4]Wayne State University, Department of Psychiatry and Behavioral Neurosciences, Detroit, Michigan, USA. {joseph.jacobson, sandra.jacobson}@wayne.edu
[5]San Diego State University, Center for Behavioral Teratology, San Diego, CA 92120, USA. {smattson, kowens, eriley}@sdsu.edu
[6]University of New Mexico, Center on Alcoholism, Substance Abuse & Addictions, Albuquerque, NM 87106, USA. pmay@unm.edu
[7]St. Vincent's Hospital, Indianapolis, IN 46032, USA. ESMoore@stvincent.org
[8]State University of New York, Buffalo, New York 14260, USA lutherkr@buffalo.edu
[9]University of Washington Medical School, Department of Psychiatry and Behavioral Sciences, Fetal Alcohol and Drug Unit, Seattle, Washington 98195, USA. astreiss@u.washington.edu.
[10]Indiana University, Office of the Vice President for Information Technology, Bloomington, IN 47405, USA. stewart@iu.edu.

For correspondence:

Andrew D. Arenson
535 W. Michigan St.
Indianapolis, IN 46202
USA

Phone: 317-278-1208
FAX: 317-278-1852
Email: aarenson@iupui.edu

Abstract

Many previous attempts by fetal alcohol spectrum disorders researchers to compare data across multiple prospective and retrospective human studies have failed due to both structural differences in the collected data as well as difficulty in coming to agreement on the precise meaning of the terminology used to describe the collected data. Although some groups of researchers have an established track record of successfully integrating data, attempts to integrate data more broadly amongst different groups of researchers have generally faltered. Lack of tools to help researchers share and integrate data has also hampered data analysis. This situation has delayed improving diagnosis, intervention, and treatment before and after birth. We worked with various researchers and research programs in the Collaborative Initiative on Fetal Alcohol Spectrum Disorders (CI-FASD) to develop a set of common data dictionaries to describe the data to be collected, including definitions of terms and specification of allowable values. The resulting data dictionaries were the basis for creating a central data repository (CI-FASD Central Repository) and software tools to input and query data. Data entry restrictions ensure that only data which conform to the data dictionaries reach the CI-FASD Central Repository. The result is an effective system for centralized and unified management of the data collected and analyzed by the initiative, including a secure, long-term data repository. CI-FASD researchers are able to integrate and analyze data of different types, collected using multiple methods, and collected from multiple populations, and data are retained for future reuse in a secure, robust repository.

Keywords: fetal alcohol spectrum disorders; data dictionary; data sharing; integrative research

Introduction

The Collaborative Initiative on Fetal Alcohol Spectrum Disorders (CI-FASD) is an international collaboration focused on finding novel methods of intervention, diagnosis, and treatment for FASD through the use of new modalities, the integration of data from animal and human studies, and the comparison of study data from multiple populations. By ‘modality’, we refer to a method of acquiring data. Modalities can differ from each other in terms of what type of data are collected, such as scores from a neuropsychological test versus images of a brain, or in terms of the method of data collection, such as measurements of facial features deriving from either a physical exam or by calculating distances on a digital, three dimensional facial model.

Many of the researchers involved in CI-FASD, though able to successfully integrate data from multiple populations with their closest collaborators, had made previous, unsuccessful attempts to share and integrate data from a larger set of populations with a larger community of FASD researchers. The FASD community is not alone in this problem, which has been noted in other research communities, such as microarray analysis (Mattes et al., 2004), where competing standards prevent data from different sites from being integrated. The problems FASD researchers encountered were multiple and varied with the different types of data that needed to be combined. Regarding

dysmorphology, for instance, pediatricians, neonatologists, and geneticists of varying experience levels used inconsistent methods to evaluate alcohol-related structural features in different subject populations. These inconsistencies and variable experience levels made it impossible to clearly define across studies by different groups which subjects had the necessary physical features to be considered for a diagnosis of Fetal Alcohol Syndrome (FAS). In terms of evaluating cognitive ability, different groups of researchers had used different instruments. When multiple neuropsychological instruments that examined a similar capability of a patient were used, the resulting measurements were still not comparable. Even when the same instrument was used in multiple populations, differences in naming variables and in the use of raw versus standardized scores led to confusion in integrating data. Equally problematic, the collection of alcohol exposure and potential confounders was implemented very differently by different groups. Integrating alcohol exposure information is particularly challenging given the need to accommodate differences in cultures. Different cultures have different norms and different ideas about what constitutes, for instance, a standard drink or a "binge" drinking pattern.

Our goal was to enable CI-FASD researchers to integrate and analyze data of different types, collected using multiple methods, and collected from multiple populations, and to retain data for future reuse in a secure, robust repository.

Materials and Methods

In order to overcome these challenges and meet the goals of comparative and integrative analysis, it was critical to establish a system based on shared language (syntactic interoperability) and shared understanding (semantic interoperability) at the outset. A shared language reduces confusion from such issues as multiple terms with the same meaning or multiple spellings of the same term. A shared understanding gives confidence that the terms being used mean the same thing to everyone. In the ideal case, every term has one meaning and every meaning has one term. If such is the case, researchers can work together (interoperate) by sharing and integrating data. Data systems that incorporate such shared languages and understanding into their programming are dubbed 'intelligent'. These needs are similar to other research communities, such as Cancer researchers who have invested large amounts of research into enabling such interoperability (Komatsoulis et al., 2008). Standards were put in place for all of the types of data to be shared in the collaboration. Slightly different methods were used to determine standards for each of the various types of data that were collected, but all of these resulted in the creation of a data dictionary that encoded the variable names, labels, definitions, and allowable values. As has been long established, data dictionaries are an essential component in "making [database] contents understandable by both users and intelligent processes" (Huff et al., 1987).

For each type of data a committee with a chairperson was given responsibility for creating a standard for what data would be collected, putting control of standards with the researchers. Different types of data required varying amounts and types of accompanying training and materials to ensure that data would be collected consistently. For example, the Dysmorphology standard relied on the expertise of a small number of experts. Those

individuals set the standard for what observations would be collected, how measurements would be made, and how subjects would be categorized (having all of the necessary physical characteristics for a diagnosis of FAS, having a significant amount but not all of the physical characteristics of FAS, not having a significant amount of the physical characteristics of FAS). These experts provided training for local pediatricians, neonatologists, and geneticists in order to promote consistent standards for dysmorphology data collection, and followed up with reviews of examinations done by the local clinicians to ensure conformance with the Dysmorphology standard.

The Neurobehavior standard involved a more collaborative approach and less in-person training, but still relied on experts engaged in a two-part process similar to that used for Dysmorphology. First, representatives from throughout the CI-FASD collaboration, with the advice of external advisors, defined both the types of measurements they would like to collect from the collaboration's subjects as well as the particular instruments that would be used to collect each type of measurement. Second, extensive training materials were created for dissemination amongst the CI-FASD members and follow-up procedures were used to ensure that tests were being administered consistently.

In contrast to the Dysmorphology and Neurobehavior data dictionaries, the creation of a controlled vocabulary standard for alcohol exposure variables, although still involving the collaboration of experts from across CI-FASD, did not necessitate the same need for extensive documentation and training. Other more novel and/or less commonly used modalities, such as 3-D facial imaging and ultrasound measurements have relied on a smaller cohort of collaborators to define the variables to be collected for CI-FASD, but still represent a standard for the consortium.

The data dictionaries collectively represent the data standards created for CI-FASD. Although each individual data dictionary generally represents a single modality, in some cases multiple modalities might be within one data dictionary or a modality might be represented in more than one data dictionary. The data in any given data dictionary were grouped together for ease of data entry purposes, as each data dictionary was used to create a single data entry application, called a data input tool.

Before developing software, it is necessary to understand what requirements the software must meet, and the data dictionaries provide documentation of these requirements. It is unrealistic, however, to expect all future needs of researchers to be known in advance. Even after due diligence has been taken to carefully consider which data are to be collected and to codify the meaning and allowable ranges for that data, situations frequently arise in which it would prove helpful for data managers to provide some commentary on the data that has been collected. Examples include noting unanticipated observations or explaining why certain data were not available. It has been shown that combining coded fields with free text fields provides the fullest assessment of a clinical database (Stein et al., 2000). Free text fields were thus provided to enhance the effectiveness of the data collected by the collaboration.

Similarly, the creation of data dictionaries proved to be an ongoing rather than onetime process. It is to be expected, as has been seen in other research (Brindis et al., 2001), that clinical practice will change over time, requiring data dictionaries, software tools, and repositories to change as well. For the most part, required changes to the data dictionaries have involved simple additions of variables or changes as to what values were allowed for a variable. At times, however, the experience of researchers attempting to capture or analyze data led to uncovering differences in interpretations of the data dictionaries that necessitated refinements or corrections. Early attempts at analysis, for instance, uncovered uncertainty amongst CI-FASD researchers as to the meaning of the FASSTATUS variable which is assigned to subjects on the basis of a physical examination, but which some researchers misunderstood to be either a diagnosis or to represent a recruitment group (i.e. subjects expected to be in one of the following categories: FASD, Control, Contrast). In response, the definition for this variable was clarified as above and an additional variable was added to separately store the recruitment group for a subject.

Although data dictionaries formalized the standards adopted by CI-FASD, and the CI-FASD Central Repository provided a place for data to be brought together for integration, sharing, and analysis, further software was required to ensure that the data collected and submitted to the Central Repository conformed to the standards set in the data dictionaries. Conformity was enforced at two points in the process of data collection: Data input tools and data submission tools.

The CI-FASD collaboration is an inherently distributed project. Not only are researchers interested in being able to collect and manage their data locally for the purposes of controlling what subset of data is submitted to the Central Repository, but in some cases lack of robust or, in fact, any Internet access absolutely required that data collection tools be separated from the Central Repository and made available at a researcher's desktop or on a laptop in the field. We provided data input tools that simultaneously allowed researchers to locally manage their data while enforcing the standards of the CI-FASD data dictionaries. These input tools were created for Dysmorphology, Neurobehavior, Alcohol & Control, and Ultrasound data. The tools are implemented as Microsoft Access® databases that provide a graphical interface for entering data and use the allowable value constraints from the data dictionaries to ensure conformance with CI-FASD standards. Applying these standards at the point of data entry provides immediate feedback to the data managers so that any nonconformance can be more easily addressed.

The second point where conformance is ensured is during the process of submitting data to the CI-FASD Central Repository. The input tools export, at the touch of a button, data in an Extensible Markup Language (XML) format that are ready to be submitted to the CI-FASD Central Repository. XML-formatted data and image files are submitted to the Central Repository through a web form that analyzes files and returns feedback about the suitability of the data being submitted. Before data are admitted to the Central Repository, the standards from the CI-FASD data dictionaries are applied to produce this feedback. Data that do not meet the standards are denied and data managers are given an opportunity to address any nonconformance and resubmit their data.

The application of conformance checks at these two points – data entry and data submission – provides both the immediate feedback needed during data entry and confidence that all data in the CI-FASD Central Repository conform to the data dictionary standards. The second conformance check at the point of submission is crucial because it is possible that data might accidentally or intentionally be altered after the point that data entry conformance checking had been applied. The second check at the data submission point also provides researchers with the flexibility to develop their own tools for data entry, as desired, while still providing a point at which to check such data for conformance to the data dictionary controlled vocabulary standards. One group followed this path in order to make use of Scantron devices for reducing the efforts of data entry, and the 3D Facial Imaging Core made use of code built into their software to generate XML-formatted data for submission to the Central Repository rather than requiring manual data entry.

Results

We have successfully provided data standards and software that enable CI-FASD researchers to combine data from disparate populations and modalities, as shown by the papers cited previously and others. The CI-FASD Central Repository allows researchers to select the data sets of interest and retrieve them in a few useful formats, including HTML for web browsing and a tab-delimited format that can be easily imported into a spreadsheet program like Excel or read using a statistical analysis package like SAS or SPSS.

Table 1 shows the modalities for which we have created data dictionaries and, where appropriate, data input software tools. The three dimensional facial imaging data were captured using a process that automatically generated data files suitable for submission to the Central Repository without the need for manual data entry.

The CI-FASD Central Repository does not include diagnoses. It includes the physical and behavioral information necessary to make a diagnosis according to standard criteria and additional information that can be used to consider alternative criteria for diagnosing FAS and FASD.

Many studies have made use of the data integration made possible by the CI-FASD data dictionaries and CI-FASD Central Repository to publish results analyzing data from multiple populations (Autti-Ramo et al., 2005), multiple modalities (Kfir et al., 2009), and often both (Moore et al., 2007).

Discussion

The practices and technologies used in developing the data dictionaries, related data entry software, and the CI-FASD Central Repository could be used for other collaborative studies. They are particularly useful in situations where researchers must find a way to

agree on a controlled vocabulary and in situations where researchers require a way to manage data locally before submitting data to a central repository.

Because a significant group of FASD researchers came to agreement on a set of data dictionaries, these data dictionaries, their associated data entry tools, and the CI-FASD Central Repository are of general use to the FASD research community and other researchers studying related health impacts of maternal alcohol consumption. Any researcher willing to transform his or her data to conform to the CI-FASD controlled vocabulary standard will be able to combine that data with data from the CI-FASD Central Repository. Just as the collaborators within CI-FASD have come together to enable larger, more varied populations and more numerous modalities to be considered when looking for better interventions, diagnoses, or treatment of FASD, by conforming to the CI-FASD data dictionaries other researchers could further expand the scale and scope of hypothesis that can be addressed.

CI-FASD has created a controlled vocabulary standard, software tools, and the persistent CI-FASD Central Repository to help meet the NIH goal of "wide access to technologies, databases, and other scientific resources that are more sensitive, more robust, and more easily adaptable to researchers' individual needs" (National Institutes of Health, 2006). We are providing the data dictionaries and data entry tools to the FASD research community to allow any FASD researcher to take advantage of the significant effort that went into creating these data standards and software tools. Efforts are ongoing within CI-FASD to enable more modalities to be shared, including new neuropsychological examinations, infant neuropsychological examinations, physiology measurements, expanded demographic measurements, nutritional variables, and possibly ultrasound videos and brain images. As data dictionaries and software tools are created for these data, they will be released to the FASD research community. Please contact the author for information about how to retrieve the latest data dictionaries and data entry tools. Currently the CI-FASD Central Repository only allows access from CI-FASD members. Please contact the author for information about how to receive data from the CI-FASD.

Acknowledgments

This research was supported primarily by the Informatics Core for the Collaborative Initiative on Fetal Alcohol Spectrum Disorders, which received support from the National Institute on Alcohol Abuse and Alcoholism under grants 1U24AA014818 and 2U24AA014818-04.

This research was supported in part by the Indiana Genomics Initiative (INGEN). The Indiana Genomics Initiative (INGEN) of Indiana University is supported in part by Lilly Endowment Inc.

This research was supported in part by Shared University Research Grants from IBM, Inc. to Indiana University.

References

1. Brindis R.G., Fitzgerald S., Anderson H.V., Shaw R.E., Weintraub W.S., Williams J.F. (2001). The American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR): building a national clinical data repository. J. Am. Coll. Cardiol. *37*, 2240-2245.
2. Moore, E.S., Ward, R.E., Flury-Wetherill, L., Rogers, J.L., Autti-Ramo, I., Fagerlund, A., Jacobson, S.W., Robinson, L.K., Hoyme, H.E., Mattson, S.N., et. al. (2007). Unique facial features distinguish fetal alcohol syndrome patients and controls in diverse ethnic populations. Alcoholism: Clinical and Experimental Research. 31(10), 1707- 1713.
3. Autti-Ramo, I., Fagerlund, A., Ervalahti, N., Loimu, L., Korkman, M., Hoyme, H.E. (2006) Fetal Alcohol Spectrum Disorders in Finland: Clinical delineation of 77 older children and adolescents. American Journal of Medical Genetics 140A: 137-143.
4. Huff, S.M., Craig, R.B., Gould, B.L., Castagno, D.L., Smilan, R.E. (1987). A Medical Data Dictionary for Decision Support Applications. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2245107
5. Kfir, M., Yevtushok, L., Onishchenko, S., Wertelecki, W., Bakhireva, L.N., Chambers, C.D., Jones, K.L., Hull, A.D. (2009). Can prenatal ultrasound detect the effects of in utero alcohol exposure? A pilot study. Ultrasound Obstet Gynecol. in press.
6. Komatsoulis, G.A., Warzel, D.B., Hartel, F.W., Shanbhag, K., Chilukuri, R., Fragoso, G., de Coronado, S., Reeves, D.M., Hadfield, J.B., Ludet, et al. (2008). caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform *41(1)*, 106-123.
7. Mattes, W.B., Pettit, S.D., Sansone, S., Bushel, P.R., Waters, M.D. (2004). Database Development in Toxicogenomics: Issues and Efforts. Environmental Health Perspectives *112*, 495-505.
8. National Institutes of Health. (2006). NIH Roadmap for medical research - Overview of the NIH Roadmap. http://nihroadmap.nih.gov/overview.asp
9. Stein H.D., Nadkarni P., Erdos J., Miller P.L. (2000). Exploring the Degree of Concordance of Coded and Textual Data in Answering Clinical Queries from a Clinical Data Repository. Am Med Inform Assoc. *7*, 42-54.

TABLE 1: Data Dictionaries and Software Tools created for the Collaborative Initiative on Fetal Alcohol Spectrum Disorders by Modality

Modality	Description	Number of Variables	Tools
Dysmorphology	Physical examination	103	Data Dictionary Input Tool
Neurobehavior	Eighteen neuropsychological tests	798	Data Dictionary Input Tool
Alcohol & Control	Maternal prospective questionnaire administered in early pregnancy	599	Data Dictionary Input Tool
3D Facial Imaging	Three-dimensional model of the face plus facial measurements	25	Data Dictionary
Ultrasound	Anatomic survey	103	Data Dictionary Input Tool