

The National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA): A Framework Supporting Neuroimaging Data Integration and Analysis

B. N. Nichols^{1, 2*}, Weiwei Chu¹ and Kilian M. Pohl^{1, 2}

¹ SRI International, Center for Health Sciences, USA

² Stanford University, Psychiatry and Behavioral Sciences, USA

Introduction

Alcohol and marijuana remain the most commonly used central nervous system-active substances in the teen years [1]. The National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) is a multisite, longitudinal “Big Data” study using quantitative assessment tools necessary to capture the influence of adolescent alcohol and marijuana abuse on neurodevelopment. To accomplish its aims, NCANDA set out to recruit 830 participants, ranging from 12-22 years old, across five data collection sites nationwide.

The NCANDA Data Analysis and Informatics Component (DAIC) facilitates electronic data capture, management, analysis, and distribution processes across five data collection sites. In [2], we described our data integration infrastructure for clinical and cognitive data that uses a distributed and version controlled approach to upload and harmonize multiple data sources (i.e. University of Pennsylvania Web-based Computerized Neurocognitive Battery (WebCNP) [3], LimeSurvey [4], Blaise [5], and ePrime [6]) to the Research Electronic Data Capture (REDCap) system. Since then, we have scaled up our approach by incorporating data and metadata from the eXtensible Neuroimaging Archive Toolkit (XNAT) with extensions to support automated data processing.

Dataset Description

Each data collection site carried out the same core assessment and sites worked in pairs to conduct additional studies (e.g., overnight sleep evaluation, recovery during monitored abstinence). The 831 study participants completed a core data acquisition protocol at baseline and will complete two annual followups including the neuropsychological (NP) test battery, neuroimaging session (MRI, DTI, and rsfMRI), a comprehensive assessment of substance use, psychiatric symptoms and diagnoses, and functioning in major life domains. In addition, a mid year phone interview is conducted between each visit to track substance use. The NP test battery assesses seven major functional domains including: general intelligence; executive functions; emotion regulation; multimodal and multiple component mnemonic processes; visuospatial abilities; basic visual acuity and color perception; and motor skills of eye-hand coordination, speed, and postural stability. Bio-samples for genetic analysis are collected annually. One parent of each youth completes an annual interview on the youth and family environment. Upon completing data collection, the dataset is expected to reach approximately 6TB of primary data and nearly 20TB of derived data from neuroimaging analyses.

Data Analysis and Informatics Platform

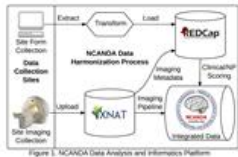
We developed a platform to automate data harmonization processes for clinical, NP tests, and brain image measures for the NCANDA study (Figure 1). All data collected using electronic data capture were automatically merged into a REDCap server hosted by the NCANDA DAIC. Data items captured on laptops at each site, rather than being directly entered into REDCap, were automatically extracted, transformed into a compliant format, and loaded into REDCap from a secure and encrypted Subversion [8] version control system. Imaging data was first uploaded from the site specific Picture Archiving and Communication Systems (PACS) to a XNAT server hosted by the NCANDA DAIC. All data were evaluated with quality control checks that included automatic test scoring, range validation, and a neuroradiologist report for incidental imaging findings. Finally, quality control processed imaging and corresponding non-imaging data for each session were harmonized within a single REDCap project to generate data integrity reports on a biweekly basis. Identified issues were resolved with site consultation for scoring irregularities, mistyped IDs, visit dates, and any data that were not been uploaded properly [2]. Data for all future manuscripts published by the consortium will be based on quality reviewed, versioned data releases provided by the NCANDA DAIC.

Conclusions and Future Work

Heterogeneous data models and semantics are inherent to complex study protocols that capture rich neuroimaging and neuropsychological measures. The diversity of data collected by these studies requires biomedical data management and electronic data capture systems tailored to specific use-cases. For example, the successful operation of the NCANDA study required the development of a data integration system to merge multi-model data from different information systems (e.g., WebCNP, REDCap,

XNAT). This is a suitable solution for use within the NCANDA consortium, but barriers remain for broader data reuse. Without a mapping to common data elements and terminologies, data sharing and integration with external resources is limited. Incorporating metadata standards in the design of this systems may streamline data integration processes and interoperability, facilitating submission to national data repositories and querying data across studies. An implementation of metadata standards is currently outside the scope of most medical imaging studies, such as NCANDA, which is a serious limitation to the longevity of collected data. Future work aims to address these shortcomings through adoption of community-driven data exchange standards, such as the Neuroimaging Data Model (NIDM, [9]).

Figure 1



Acknowledgements

We acknowledge the efforts of the NCANDA Consortium's data analysis and informatics core (5U01AA021697), administrative core (5U01AA021695), and data collection sites (5U01AA021692, 5U01AA021696, 5U01AA021690, 5U01AA021681, 5U01AA021691) through the support of U.S. National Institute on Alcohol Abuse and Alcoholism with co-funding from the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute Child Health and Human Development.

References

1. L. D. Johnston, P. M. O'Malley, R. A. Miech, J. G. Bachman, and J. E. Schulenberg, "Monitoring the Future national survey results on drug use: 1975-2014: Overview, key findings on adolescent drug use." 2015.
2. T. Rohlfing, K. Cummins, T. Henthorn, W. Chu, and B. N. Nichols, "N-CANDA data integration: anatomy of an asynchronous infrastructure for multi-site, multi-instrument longitudinal data capture," *Journal of the American Medical Informatics Association*, pp. amiajnl-2013-002367, 2013.
3. WebCNP: <https://webcnp.med.upenn.edu/>
4. LimeSurvey: <http://www.limesurvey.org/>
5. Blaise: <http://www.blaise.com>
6. ePrime: <http://www.pstnet.com/eprime.cfm>
7. Marcus, D., Olsen, T., Ramaratnam, M., & Buckner, R. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics*, 5(1), 11-33.
8. Subversion: <https://subversion.apache.org/>
9. Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *NeuroImage*, 82, 647-661. <http://doi.org/10.1016/j.neuroimage.2013.05.094>

Keywords: neuroinformatics, data integration, data analysis, Neuroimaging, data sharing, Addiction, alcohol use disorder

Conference: Neuroinformatics 2015, Cairns, Australia, 20 Aug - 22 Aug, 2015. **Presentation Type:** Poster, to be considered for oral presentation **Topic:** Neuroimaging

Citation: Nichols BN, Chu W and Pohl KM (2015). The National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA): A Framework Supporting Neuroimaging Data Integration and Analysis. *Front. Neurosci. Conference Abstract: Neuroinformatics 2015*. doi: 10.3389/conf.fnins.2015.91.00042

Received: 02 May 2015; **Published Online:** 05 Aug 2015.

* **Correspondence:** PhD. B. N. Nichols, SRI International, Center for Health Sciences, Menlo Park, WA, USA, nolan.nichols@gmail.com