Funded by NIMH R01MH116156-01A1 (PI: Nielson)

Overview of Repository Search Strategy

The NIMH Data Archive (NDA) and Federal Interagency Traumatic Brain Injury Research (FITBIR) repositories share numerous assessment domains in-common yet are not presently interoperable. Therefore, we developed a system for matching the data structures and the specific elements included within those structures across repositories. This system has been used to align the databases using three levels of precision, as displayed in the Alignment Process Diagram on page 3. The broadest level of precision involves aligning the overall data structures that represent entire assessments or cognitive batteries by name and generation (e.g., first edition, second edition, etc.). This first step provides an overview of what potential assessment domains are shared across repositories based on name similarity. The next level of precision involves aligning the specific elements within each assessment that correspond to unique items or assessment scores, using a combination of the names of items and the research team's prior domain knowledge of the assessment instruments. This second step allows us to determine what components of each assessment are included in both batteries, and which can potentially be aligned. Finally, the third level of precision involves examining the scaling used for each item or assessment score to determine how these elements align across the data repositories. This final step allows us to determine the extent of transformation needed to match elements across repositories.

Summary of Search Results

We have presently confirmed alignment of 161 of FITBIR's 711 data structures (22.6% of structures represented in the entire repository) with 122 of NDA's 3140 data structures (3.9% of structures in the entire repository). Within those structures, we have confirmed alignment of 1971 elements across 87 distinct assessment domains (e.g., separate IQ tests, diagnostic assessments, TBI evaluations, etc). Of these elements, 1059 matched identically across the repositories, meaning that no transformation is needed to pair these elements. Another 140 elements showed equivalence in scaling and only required additive or subtractive transformations to be aligned. For instance, an item with a lower anchor of 0 and upper anchor of 4 in NDA may need to be transformed through the addition of 1 to each score to achieve identical alignment with the same FITBIR item scaled from 1 to 5.

We identified a third group of 312 elements requiring other transformations or additional information included in the data structures to be aligned across databases. In some instances, these elements required multiplication, division, rescaling, or reverse-coding transformations to be matched across databases. For other elements in this category, information from other variables would be needed to align across NDA and FITBIR repositories. For instance, certain cognitive assessments in FITBIR use one variable to designate the trial or item type in the assessment and another to designate the participant's score on that trial or item. In such cases, both the trial/item type and the participant score variables are needed to correctly align with the corresponding NDA elements.

The fourth and final group of elements reflected items or scale scores that were aligned by name but were either differentially scored or had insufficient information about scaling to be

aligned based on the data dictionaries in either repository. Such elements are unable to be matched without further information about what scaling approach was used in one or both repositories. This information may be gathered in the future through examining the item and scale values included in the downloaded data or contacting the research teams who may have information on the original scaling of the measures.

Future Directions

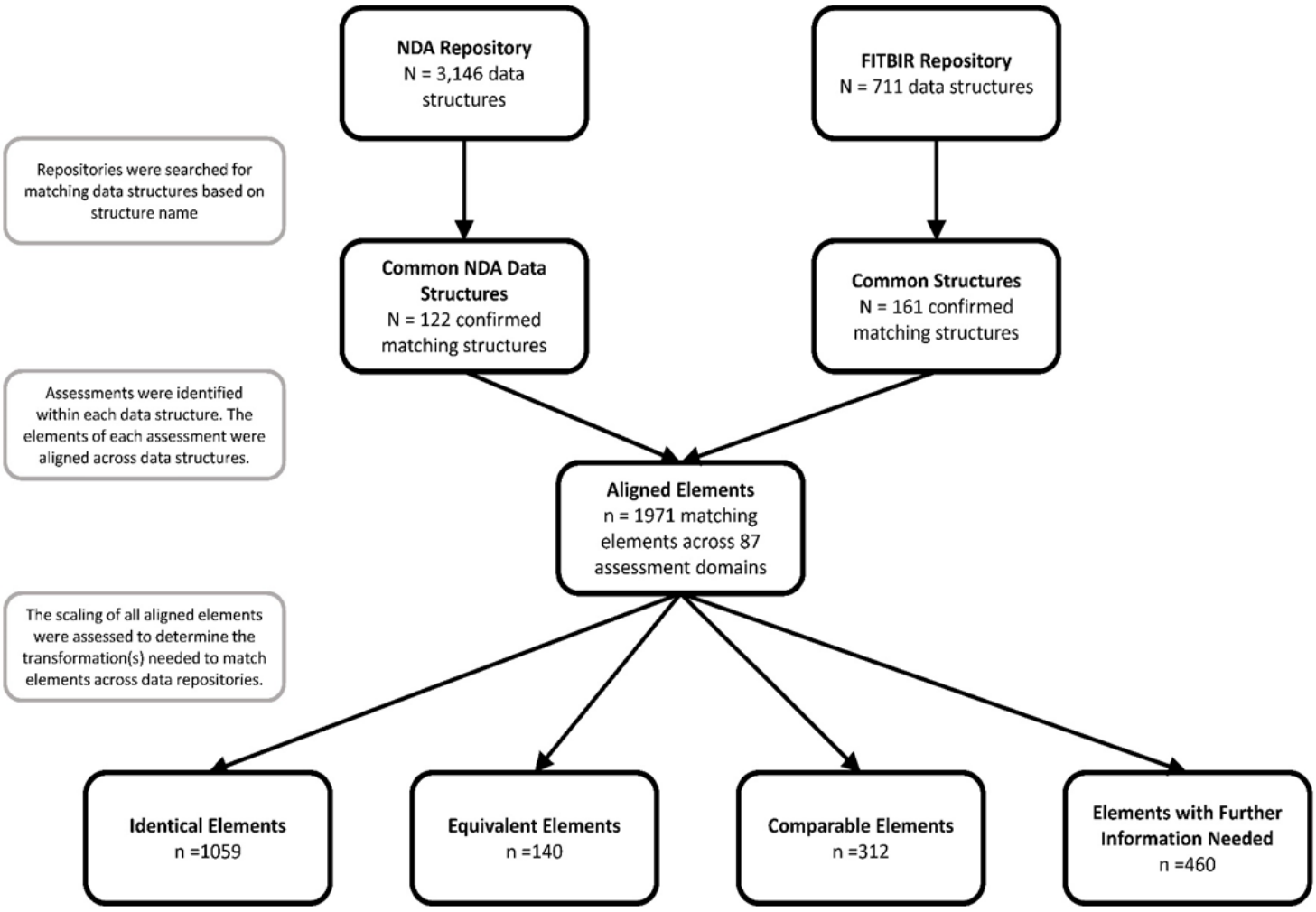
Future steps in aligning the data repositories include performing a more detailed, domain-informed search of data structures and further investigation into matching data elements that may appear in incongruent data structures. We anticipate many further data structures and elements will be aligned based on the following strategies.

The present search process has largely been guided based on alignment between the names of data structures in either repository. While this has produced a substantial number of initial results, additional matches may be derived through searching each data repository by domain (e.g., "pediatric cognitive assessments") versus scale names (e.g., "Wechsler Preschool and Primary Scale of Intelligence"), as well as further probing data structures with vague or broad names (e.g., the "Depression Survey" in NDA) to determine the inclusion of validated assessments that could be matched across repositories. We have already observed some assessments with substantially different titles across repositories, and expect this additional search to be fruitful for identifying further data structures for alignment.

At the level of data elements (e.g., specific items or scores), we plan to investigate elements that appear in incongruent data structures yet refer to the same score or values. For instance, several cognitive batteries include common tasks (e.g., the serial 7's task or the stroop task) that may be equivalently scaled across differentially-named batteries. Similarly, diagnostic screening tests often include equivalent questions if they are based on a similar diagnostic nosology (e.g., questions asking about DSM-V, Criterion A events for a PTSD diagnosis). As such, we expect it will be beneficial to further probe into element-wise alignment outside of shared data structures across the databases.

Ultimately, we aim to incorporate the results of our present matching system into the NDA-FITBIR Pipeline application we have developed for aligning data within each repository. This will allow researchers using the app to maximize potential measures and participants drawn from each repository, and will lay a foundation for a future user-friendly integration of the NDA and FITBIR databases.

NDA-FITBIR Alignment Process Diagram.



Common structures = Data structures (i.e., forms containing multiple items or scores) with the same or similar names across databases. Aligned Elements = Items or scores with identical names or semantics across databases. Identical Elements = Elements with identical scaling across databases. Equivalent Elements = Elements with the same range of scores that require only addition or subtraction to be converted across databases. Comparable Elements = Elements of the same measure that require transformations beyond addition or subtraction across databases. Elements with Further Information Needed = Elements with identical names or semantics but distinct scaling or a lack of information on scaling, such that they cannot be aligned based on information in the data dictionaries alone.