The CRI Knowledge Network
NATIONAL DATA WAREHOUSE
BACKGROUND

Centerstone Research Institute (CRI) is a private, not-for-profit research institute created as part of Centerstone, one of the nation’s largest providers of community-based behavioral healthcare. CRI strives to connect the science and service of mental health through information technology, public policy advocacy, data analytics, program evaluation, and clinical research support. Although mental health and neuroscience comprise a rapidly growing sector of medical research worldwide, there are significant delays in actually realizing the fruits of that research in clinical practice. CRI brings together academics, technologists, clinicians, industry leaders, and philanthropists to drive fundamental advancement in the practice of mental and behavioral health.

In 2008, CRI initiated a partnership with key behavioral health providers beyond the Centerstone system, researchers, and policymakers nationwide to launch the Knowledge Network — a collaborative organization dedicated to bridging the gap between mental health research and practice. Knowledge Network members work together to drive implementation of evidence-based treatments from the best research available, ultimately improving the lives and outcomes of patients nationwide. The Knowledge Network also seeks to increase the knowledge base on effective practices by leveraging their combined data in pursuit of practice-based evidence. Over its lifetime, CRI’s Knowledge Network has launched a series of projects designed to advance research, policy, treatment, and technology in the mental health arena.

“To improve clinical practice as a whole and then evaluate it, healthcare must go beyond basic automation and convert electronic patient record data into aggregated, multidimensional information”, argue Ledbetter and Morgan in support of clinical data warehousing. The Knowledge Network is taking a significant step in this direction for mental and behavioral health with its National Data Warehouse project. Through the project, participating community mental health centers across the country have contributed practice data to a central repository, resulting in what is today the largest community mental health data store of its kind.

The National Data Warehouse now contains data from 14 organizations, covering over 1,100 separate clinical practice locations across 13 states. Details are included on more than 480,000 patients, 23 million service records, 4.5 million prescriptions, 1.9 million diagnoses, and a variety of other clinical, administrative, and outcomes data points related to the practice of mental and behavioral health. Over 100,000 incoming files have been processed into the database since its inception.

This paper will describe the National Data Warehouse and its contents, provide examples of the kinds of investigation the dataset could support, and discuss opportunities for other organizations to participate in the project. It is the goal of CRI’s Knowledge Network that the National Data Warehouse continues to mature into a broad and useful mental health data resource.
**METHODOLOGY**

The National Data Warehouse was conceived by the cross-organizational Technology Workgroup of the Knowledge Network. Collaborators from major community mental health centers and research institutions defined a standard set of data elements to make up the core of the dataset, allowing for inclusion of as much or as little data as the contributing organization is able to provide. The data model, database, and related programs were then implemented by the analytics department of CRI, which continues to maintain and operate the system today.

The National Data Warehouse is a continuously-growing data store, and has been in operation since 2011. Files are transmitted securely from the electronic health record systems of participating organizations to CRI in a standardized file format for processing. Each night, CRI executes an automated process to further standardize and load available data. Information from all organizations is then combined into a single relational data model hosted by CRI, and the organization-specific records are further enriched with standard clinical dictionaries, NDC drug categorization, geographic coding, weather patterns, and other external data.

The CRI team responsible for implementation of the system was recognized in 2010 as a Best Practices winner by The Data Warehousing Institute (TDWI) ⁵. The data model used by the National Data Warehouse has also found success as part of CRI’s Enlighten Analytics⁶ package of analytical tools and reports, currently marketed in the commercial space by a leading behavioral health information technology vendor.

**CONTENT AND APPLICATIONS**

The source organizations are community mental health centers, and the population represented is approximately 90% funded by Medicaid, grants, and other safety net mental health programs and agencies. A small proportion of the services are covered by commercial insurance, self-pay, or other financial arrangements. The project actively began accepting data in 2011, with additional organizations rolling on over time. Partners were asked to back-populate data in the warehouse to January 2009 if available in their electronic health record systems.

The core dataset, conceptualized in Figure 1, is centered on patients and services. Demographic information includes gender, age, race, and ethnicity as available from source systems. Each service record includes the timing and location of the service, the type of activity included, payor and (optional) revenue information, CPT codes, staff credentials, organization-specific program information, appointment status (including missed appointments), and a variety of other service-related fields. Details on relevant primary and secondary diagnoses for each patient across DSM-IV Axis I and Axis II, including diagnosis code and timing, are also included as available.

Some participating organizations include additional information. For example, prescription details are often available, which is then matched to FDA NDC reference data for standardization. For a subset of contributing organizations, a limited set of outcomes data is available: the SAMHSA National Outcomes Measures (NOMs)⁷, the Adult Needs and Strengths Assessment (ANSA)⁸, and Childhood and Adolescent Needs and Strengths (CANS)⁹. There is also potential within the data to calculate administrative outcomes (e.g. level of patient engagement and length of treatment / episode of care).

Since the National Data Warehouse is compiled directly from electronic health records as opposed to a secondary data source like insurance claims, it contains information that would never make it to a claim — for example, appointments that were scheduled but never kept, and services that were rendered pro-bono and never billed. This also makes it feasible to begin analysis of a new program or policy without waiting for claims processing to occur.

Internally, CRI is leveraging the Centerstone portion of the dataset for analytics around service management, financial management, staff productivity, quality, and outcomes reporting. The CRI analytics department has also implemented predictive models for Centerstone revenue projection and patient-level hospitalization prediction. That said, CRI believes this data resource contains broad research potential well beyond what has been pursued so far.

The dataset has potential to be used for detailed investigation on a variety of questions:

- How does behavioral health service delivery differ across different payors or geographies?
- How accurate is the initial diagnosis? How do diagnoses change over time?
- What influences patients to engage (or not engage) with care?
- What are the impacts of different prescribing patterns, 1st/2nd-line drug treatments, etc.?
- How do the highest-utilizing patients differ from the rest of the population?
- What are the hidden costs of pro-bono / unbilled services?
- How do multiple comorbid diagnoses affect care?
CRI is in the process of preparing and certifying a de-identified set of extracts from the dataset, compliant with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, for public consumption. The public dataset will lose some resolution in order to meet HIPAA Safe Harbor provisions – for example, patient address details and exact service dates will not be included – but the resulting data will be a rich source of community mental health practice detail not commonly available from claims or other public sources.

**EXPLORATORY ANALYSES**

This section presents a number of analytical examples built from National Data Warehouse content.

These examples have not been thoroughly vetted for statistical significance – they simply serve to illustrate the potential of the data contained in the dataset. This analysis is focused on the period of 2011-2014, and unbroken periods of care are grouped into episodes (where 90 days with no service marks the end of an episode). The total number of episodes included in this subset is 437,151, representing 326,653 unique patients.

For measures referencing engaged patients, “engagement” means the episode included at least four services. Episodes have been grouped by diagnosis, based on diagnoses occurring in the dataset during or before the episode of care. These groups are not mutually exclusive: episodes are included in multiple diagnosis groups for those patients with multiple diagnoses.

In Figure 2, the patient age distribution shows a clear dip from age 18 into the mid-to-late twenties—suggesting, for example, that access to care or the desire to engage in treatment may be limited for young adults. Also notable is that childhood mental health patients in the dataset are clearly weighted toward males, while adulthood patients are majority female.

Figure 3 depicts a high-level distribution of kept services by self-reported activity classes in the database. Each contributing organization groups its services into a standardized set of classes based on shared definitions.

Patients are grouped by diagnosis for the next set of analyses. Figure 4 shows how age and gender patterns differ by diagnosis – for example, ADHD reflects predominantly male children, while patients being treated for depressive disorders are more typically middle-age females. Figure 5 shows variation in the...
length of episodes of care for the different diagnosis groupings, with substance abuse tending toward the shortest duration per episode.

Taking the same analysis a step further for psychoticism patients, Figure 6 shows that episodes of care for this subset of the population tend to lengthen as patients age. This is potentially indicative of chronically psychotic patients trending upward in treatment adherence over time.

Figure 7 calculates distinct medications per episode (collapsing the various forms and brands of each medication into one), and shows the average value for each diagnosis group. There is an interesting degree of variability here, with bipolar and psychoticism patients receiving on average the largest number of distinct medications for each episode of care.

The final set of analyses features kept appointments (the percentage of appointments made that were actually attended). Figure 8 (pg 6) shows variability in the average kept rate for patients across the diagnosis groupings. Note that this only includes engaged patients — when disengaged patients and initial no-shows are included, the average kept rate is lower. Figure 9 (pg 6) shows another interesting phenomenon for chronic patients: As episode length increases, kept rates increase as well, perhaps as patients becomes more comfortable with treatment.

**LIMITATIONS**

The Knowledge Network National Data Warehouse has certain limitations which must be taken into account for any investigation. Obviously the available data do not provide a complete picture of a patient’s health: the dataset only contains information from community mental health centers, and does not include hospitalizations or physical health information. It also does not currently provide extensive clinical depth, as details about the specifics of mental health treatment are typically hidden in clinical notes or other unstructured data stores. The dataset also cannot be considered a full longitudinal view of a patient’s mental health services as it only captures events occurring at participating organizations during the time period covered by the project.

Standardization of data across organizations is a continual challenge in mental healthcare. The National Data Warehouse project attempts to apply some standardization to allow for cross-organizational study, but there are still
significant challenges that must be considered as part of any investigation. CPT and other coding mechanisms are not used in a standard way across the industry, and there is no standard method of classifying activities. The National Data Warehouse allows organizations to self-map activities into Activity Classes (therapy, psychosocial, case management, etc.) but these mappings are not necessarily stable over time or across organizations.

Timing of available data differs by contributing organization; some organizations include data for only a limited timeline (one or two years in some cases), due either to local resource availability or the implementation timeline of their electronic health record. Medication and outcomes assessment data are only available from a subset of contributing organizations. CRI continues to expand and broaden the dataset as new data become available and as new organizations get involved in the project.

GETTING INVOLVED

The National Data Warehouse continues to mature and grow. CRI is actively seeking partnerships with organizations interested in using the data for scientific advancement and policy research, or who can contribute to and further enrich this shared data asset.

For researchers interested in exploring the use of this dataset to further your research: CRI can work with you to discuss goals and perform preliminary analysis. Once it’s determined the National Data Warehouse’s capabilities are aligned with your research goals, CRI can collaborate with you on the best way to proceed given your project’s goals and funding. CRI staff may be available for query and analytical work, or it may be possible to establish direct access for your team under appropriate agreements. CRI is also working toward a limited, de-identified public dataset which will open a subset of the data to the global audience.

For organizations interested in contributing data to the National Data Warehouse: Your organization’s participation in the National Data Warehouse project will directly help advance the science and public policy of mental and behavioral health. This is a unique opportunity to impact the future of the field with minimal resource commitment: CRI will work with you to identify the best method to extract data from your electronic health record into standard file formats, and CRI’s
expertise in data handling will ensure security and HIPAA compliance throughout the process.

To discuss how you and your organization can get involved with CRI’s Knowledge Network and the National Data Warehouse, please contact the CRI analytics department:

Centerstone Research Institute
Analytics Department
44 Vantage Way, Suite 280
615-463-6240
analytics@centerstone.org

CONCLUSION

CRI is dedicated to the prevention and cure of mental illness, and the Knowledge Network National Data Warehouse is a broad, collaborative mental and behavioral health data resource that can help further these goals. This paper discussed the content and methodology of the National Data Warehouse, surveyed examples of applications of the data, and presented several scenarios to stimulate interest in potential further investigation. The intent of CRI’s Knowledge Network and this paper is that the National Data Warehouse will open the door for further investigation of the practice of community mental health, and that the dataset will continue to mature as a rich, collaborative data resource into the future.

CRI would like to thank the Ayers Foundation and the Joe C. Davis Foundation for their generous support of this project, as well as Knowledge Network member organizations for contributions of data and expertise.

REFERENCES


You may share, copy and redistribute this material in any medium or format for any purpose under the Creative Commons license. You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material.