



BRICS

Biomedical Research
Informatics Computing System

Addressing Gaps, Challenges, and Opportunities Related to Metadata Standards

Moving the Field Forward with Technology: A Data Scientist's Perspective

Maria Bagonis, PhD
NIDDK Data and Metadata Standards Workshop
June 4th, 2025



Defining the Gaps: A Personal Research Story

HUGE disconnect between:

A Predictive AI Model, good enough for a Nature paper

VS

A Predictive AI Model that can be productionized and released in the clinic.

**Clinical
Predictive AI Model**



Lots of High-Quality DATA!



**Good Meta Data Standards;
Community Cooperation**

**Academic
Predictive AI Model**

**No Meta Data Standards;
Isolated Labs**

Reproducing/Extending a Model Requires Information Regarding:

- *Sampling Distributions of the Training Set vs Test Set*
- *Exact Feature Extraction Processing Pipeline*
- *The Final Model (Architecture and any Parameter Optimization)*



Large, Well-documented, Multi-site Datasets are needed for Biomarker Development and Clinical Translation of Predictive Models To:

- Overcome Small Sample Size Sampling Bias
- Embrace Patient Heterogeneity

Given the magnitude of these tasks, this work CANNOT be realistically achieved by a single lab or even one consortium. It requires a committed community working together effectively as a team.

Led to the Establishment of NIH-funded centralized Repositories:

BRICS



PDBP



eyeGENE
A program of the National Institutes of Health

NIMH



NDA

Includes Featured Consortium Datasets

FEATURED DATASET

**Adolescent Brain Cognitive
Development Study (ABCD)**

Use Featured Datasets to quickly access commonly requested datasets

But how do we efficiently query, aggregate data, and synthesize information from these repositories to optimize return on investment?

Challenges: What Makes a Common Data Element 'Common'?

A set of quality CDEs enables a researcher to **quickly and systematically** identify datasets that can be combined and/or compared.

The Logistically Challenging Common Data Element (CDE) Paradox:

CDEs need to be:

- Structured enough to ensure that the necessary information is stored, standardized whenever possible, and communicated effectively, to facilitate dataset aggregation and comparison.

i.e. Eliminate Unintentional, Unnecessary Divergences

- Flexible enough to evolve with changing technology and goals:
i.e., Keep Intentional, Scientifically Necessary Divergences;

How do we deal with the nuances of this paradox on a large scale?

Opportunities: The AI Large Language Model Revolution

Text Embedding Models and Generative Large Language Models Make it Easier for Computers to Help us Synthesize Detailed, Nuanced Text Quickly, Interactively, and Automatically

Current AI Projects At BRICS to:

- **Cross-map and harmonize new measurement variables of interest with existing common data element dictionaries**, thereby reducing the introduction of unnecessary redundancy, facilitating automated data aggregation, and ensuring robust clinical data standards that meet interoperability requirements.
- **Automatically identify and aggregate relevant cohort-specific datasets across studies and/or repositories** that can be combined or compared to answer a specific research question, facilitating machine learning and AI applications, as well as efficient biomarker development.

Avoiding the Data Element (DE) Dictionary Explosion: *Reducing Unintentional Redundancy*



BRICS Tool Development: Approaches to Facilitate Data Dictionary Searches



UMLS Code Assignments to Group Common Data Elements

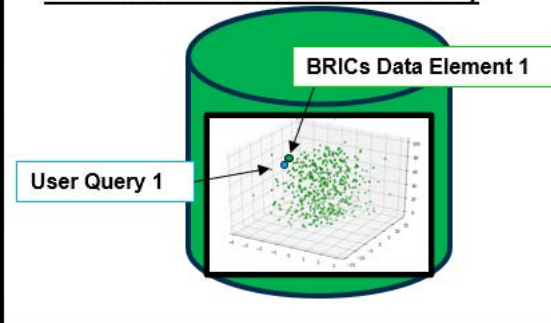
(Standardized Concept Assignments to DEs
Via Manual Curation)

UMLS Concept1:

BRICS Data Element 1
BRICS Data Element 2
BRICS Data Element 3...
Etc

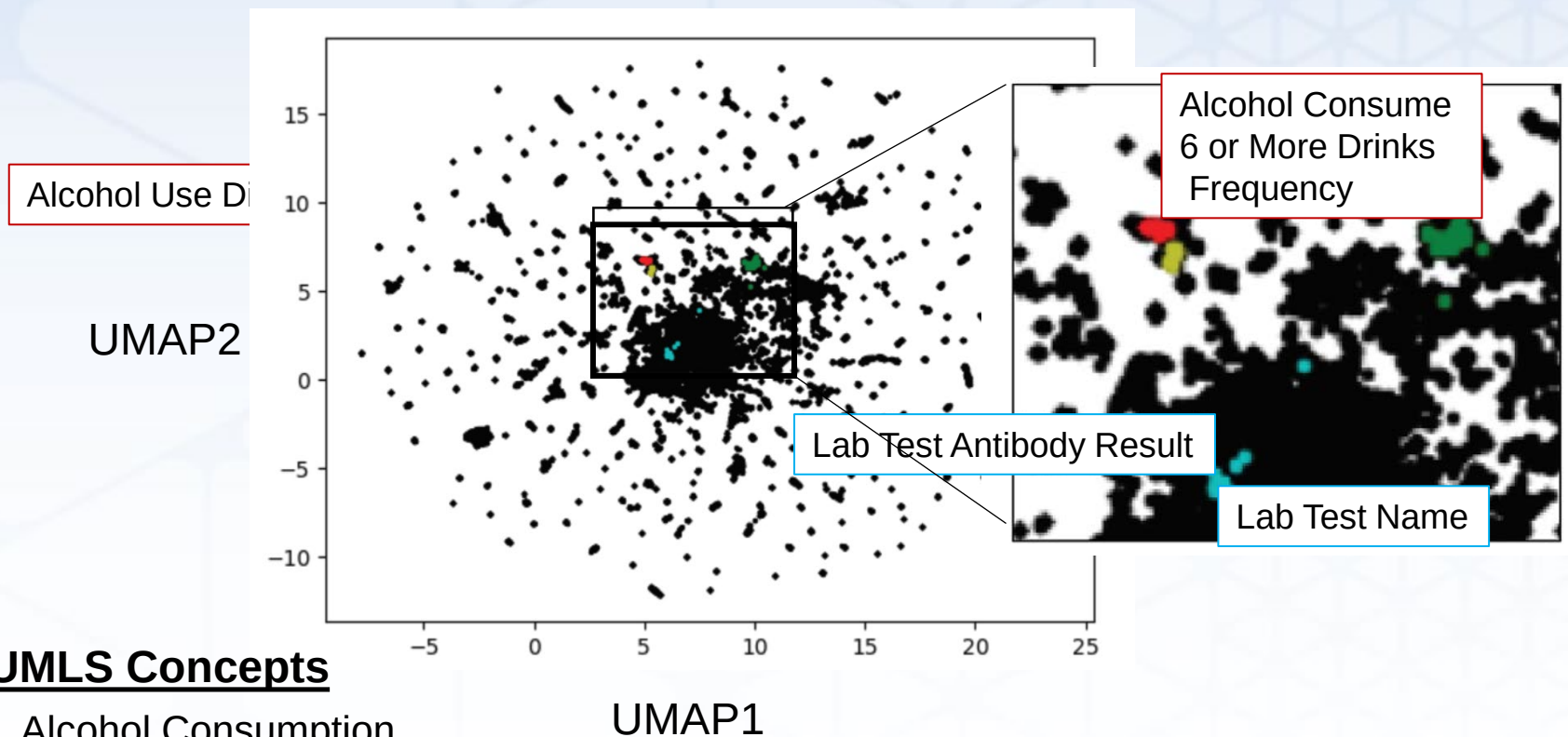
**Semantic Similarity Searches
to Group Common Data Elements**
Pre-Trained Large
Language Models transform DE text
such that DEs with similar meaning
cluster in embedding space

BRICS Embedded Vector Database: Current BRICS Data Element Dictionary



Text Embedding Models Automatically Cluster Common Data Elements by Concept

LLM Embedding Space: FITBIR Data Dictionary (Vector = 768)



UMLS Concepts

- Alcohol Consumption
- Illicit Drugs
- Laboratory Procedures
- Headache

Specific AI Projects in Development at BRICs: Data Element Dictionary Cross-Mapping Tool

AIMS4DE: Artificial Intelligence Meaningful Search for Data Elements

Public Beta-Version: <https://brics.cit.nih.gov/aims4de/>

Currently, searches FITBIR data element dictionary only, but an update allowing searches over all BRICs Instances and additional features will be pushed to the public soon.

Query Results:

☐ Filter Results By Permissible Value Descriptions

Top 5 data element matches for your query: 'traffic accident due to substance abuse' with a similarity score greater than 0.5

- Click on the left-most column of the BRICS data element (row) that best matches the text description

	Similarity Metric	Results: variable name	Results: element type	Results: title	Results: definition
☒	0.6336	TrffAccdntSelfDrngInflkhdTyp	Common Data Element	Traffic accident self drug influence likelihood type	Likelihood of whether the participant
	0.6173	TrffAccdntSelfAlcnInflkhdTyp	Common Data Element	Traffic accident self alcohol influence likelihood type	Likelihood of whether the participant
	0.5681	TrffAccdntOthPrtyDrngInflkhdTyp	Common Data Element	Traffic accident other party drug influence likelihood type	Likelihood of whether the other invc
	0.5441	SubUseInd	Common Data Element	Substance use indicator	Indicate whether or not the subject s
	0.5394	TrffAccdntOthPrtyAlcnInflkhdTyp	Common Data Element	Traffic accident other party alcohol influence likelihood type	Likelihood of whether the other invc

Performs well where traditional keyword searches fail

Incorporating the same semantic search method into different UI/Tools:

- [Data Mapping and Transformation \(DMT\) Tool](#) (IP)
- [Data Discovery and Dataset Visualization](#): (IP)

Form Structure: Collection of Common Data Elements

Form Structure: Neurological Assessment: Glasgow Coma Scale (GCS) and Pupils

This form structure is an organized set of data definitions for a form that has not been copyrighted

Form Details | **Data Elements** | Documentation | Keywords

Logically grouped data elements with defined frequency at which they repeat

#	TITLE	SHORT DESCRIPTION	VARIABLE NAME	REQUIRED?	TYPE
1	GUID...	Global Unique ID which uniquely identifies a subject	GUID	Required	CDE

... ..

GLASGOW COMA SCALE (REPEAT INFINITELY)

#	TITLE	SHORT DESCRIPTION	VARIABLE NAME	REQUIRED?	TYPE
1	Glasgow Coma Scale (GCS) - Confounders type	Type of confounders influencing participant's/subject's Glasgow Coma Scale (GCS) score	GCSConfounderTyp	Recommended	CDE
2	Glasgow Coma Scale (GCS) - eye response scale	Glasgow Coma Scale (GCS) - Best eye response (E). The GCS is a standardized instrument for assessing the level of consciousness. It evaluates three aspects of responsiveness: eye opening, motor response, verbal response.	GCSEyeRespsScale	Recommended	CDE
3	Glasgow Coma Scale (GCS) - motor response scale	Glasgow Coma Scale (GCS) - Best motor response (M). The GCS is a standardized instrument for assessing the level of consciousness. It evaluates three aspects of responsiveness: eye opening, motor response, verbal response.	GCSMotorRespsScale	Recommended	CDE
4	Glasgow Coma Scale (GCS) - verbal response scale	Score that describes the participant's verbal response according to the Glasgow Coma Scale (GCS)	GCSVerbalRespsScale	Recommended	CDE

... ..

User searches forms via a simple keyword search

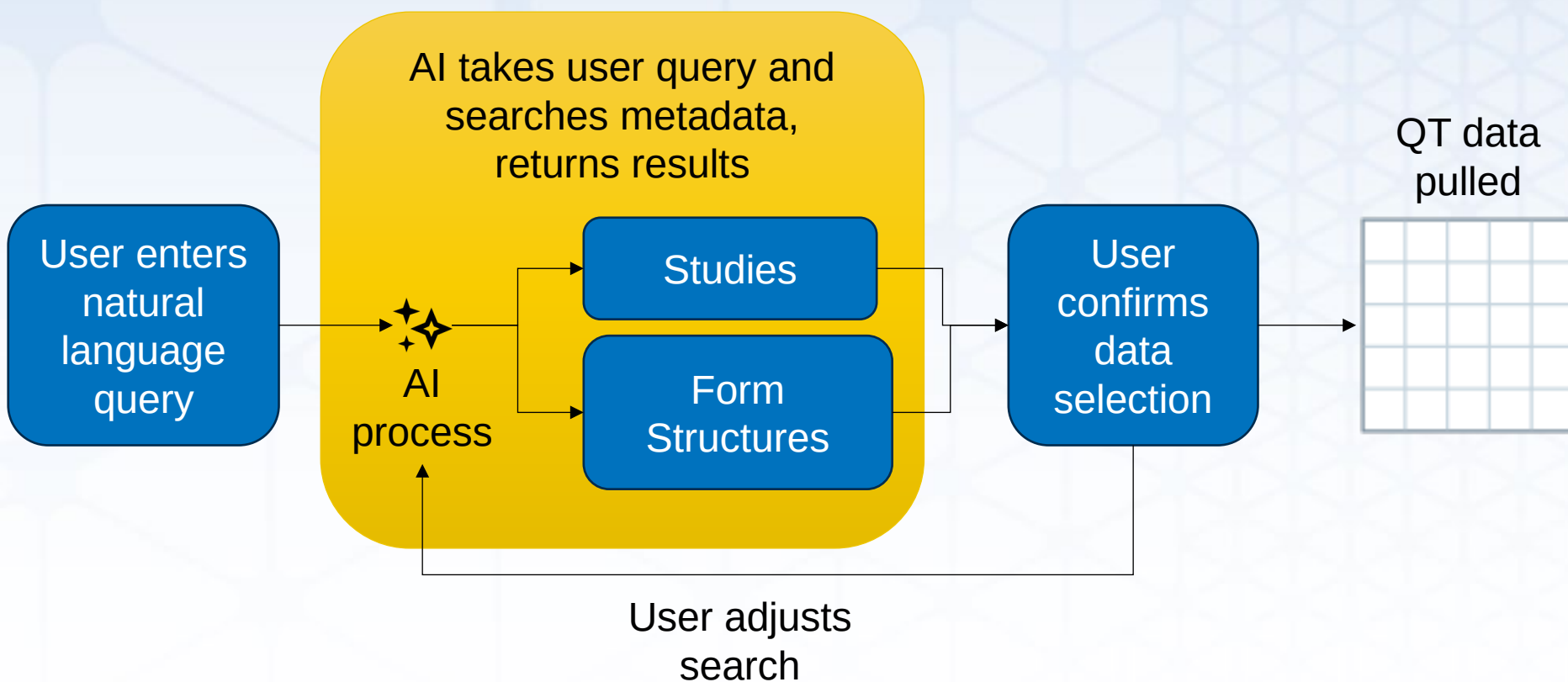
Data from all studies using the same **Form Structure** is compiled

User confirms data selection: Input additional cohort filters if desired (ie, Filter by Age, Time of Measurement, etc)

QT data pulled

- Can be difficult to identify combinations of forms of interest,
- Tedious/Time-consuming to construct cohort filters,
- No way to quickly check for sampling distributions in the aggregated dataset, without downloading

Specific AI Projects in Development at BRICs: AI for Dataset Queries





Specific AI Projects in Development at BRICs: AI for Dataset Queries

The screenshot displays the FITBIR Informatics System interface. The top navigation bar includes links for Home, ProFoRMS, Subject Management, Data Dictionary, Data Repository, Query (active), Meta Study, Account Management, and Forum. A user welcome message and a 'Log Out' link are in the top right. Below the navigation bar, there's a search section with 'Search with AI' and 'Select Data' buttons. The main content area is titled 'Search with AI using Natural Language' and shows a search query: 'A study with at least 100 participants that uses a GOS-E form structure'. Below the query, it says 'Here's what I found:' and lists '11 Studies (2 top matches)' and '6 Forms (1 top match)'. A 'Learn more about these results' button is present. A feedback section asks 'Did we find what you were looking for?' with 'Yes' and 'No' options. The search results are displayed in a grid, with 'Top Matches' highlighted. The 'Top Matches' section includes two results: 'AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury' and 'Adding Legacy Clinical Data to the Federal Interagency Traumatic Brain Injury: Citoline Brain Injury Treatment Trial'. The 'Other Results' section lists several other studies, including 'Biomarker-Based Precision Medicine Approach to Traumatic Brain Injury Subphenotypes', 'CENC Study 1: Observational Study on Late Neurologic Effects of OEF OIF OND Combat', 'Evaluation of Longitudinal outcomes in mild TBI Active-Duty Military and Veterans', 'Hyberbaric Oxygen Brain Injury Treatment (HOBIT) Trial', 'LIMBIC-CENC Prospective Longitudinal Study', 'Managing Severe TBI without ICP Monitoring: Guidelines Development and Testing (CBG Study)', 'Progesterone for the Treatment of Traumatic Brain Injury (ProTECT III)', 'Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) - Adult', and 'AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury'.

Search with AI | Select Data

Admin Only: [Clear Cache](#)

Search with AI using Natural Language

What are you looking for?

Describe what you're looking for within form structures, studies, and data elements, and provide details around specific data you'd like to see. Please check the results for accuracy.

A study with at least 100 participants that uses a GOS-E form structure

Here's what I found:

- 11 Studies (2 top matches)
- 6 Forms (1 top match)

[Learn more about these results](#)

Select the relevant forms and studies and click Select Data.

Did we find what you were looking for? ☒ Yes ☐ No

[Search with AI](#) [Clear Search](#)

Studies | **Forms**

Search: Search Studies Data Types: Data Type - All

☐ Show Only My Studies ☒ Show Accessible Studies ☐ Show All Results: 11 Studies

Top Matches: [Select All Top Matches](#)

- ☒ **AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury**
PI: Hilaire J Thompson 20 272
[Why is this a top match?](#)
- ☒ **Adding Legacy Clinical Data to the Federal Interagency Traumatic Brain Injury: Citoline Brain Injury Treatment Trial**
PI: Ross Zafonte, DO 56 1213
[Why is this a top match?](#)

Other Results:

- ☒ **Biomarker-Based Precision Medicine Approach to Traumatic Brain Injury Subphenotypes**
PI: Richard Rubenstein 46 346
[Why is this a match?](#)
- ☒ **CENC Study 1: Observational Study on Late Neurologic Effects of OEF OIF OND Combat**
PI: William Walker, MD 94 1615
[Why is this a match?](#)
- ☒ **Evaluation of Longitudinal outcomes in mild TBI Active-Duty Military and Veterans**
PI: Christine M MacDonald, PhD 46 346
[Why is this a match?](#)
- ☒ **Hyberbaric Oxygen Brain Injury Treatment (HOBIT) Trial**
PI: Gaylan L Rockswold, MD 94 1615
[Why is this a match?](#)
- ☒ **LIMBIC-CENC Prospective Longitudinal Study**
PI: William Walker 23 257
[Why is this a match?](#)
- ☒ **Managing Severe TBI without ICP Monitoring: Guidelines Development and Testing (CBG Study)**
PI: Randall Chestnut, MD, FCCM 12 111
[Why is this a match?](#)
- ☒ **Progesterone for the Treatment of Traumatic Brain Injury (ProTECT III)**
PI: David Wright 94 3628
[Why is this a match?](#)
- ☒ **Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) - Adult**
PI: Geoffrey Manley, MD, PhD 14 927
[Why is this a match?](#)
- ☒ **AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury**
PI: Hilaire J Thompson 39 882
[Why is this a match?](#)

AI shows results, sorted into Studies & Forms, and provides summary of results

User has opportunity to provide feedback

Top matches highlighted, with one click ability to select

Other results listed below



Specific AI Projects in Development at BRICs: AI for Dataset Queries

Expandable accordion to view full summary

Search with AI using Natural Language

What are you looking for?

Describe what you're looking for within form structures, studies, and data elements, and provide details around specific data you'd like to see. Please check the results for accuracy.

A study with at least 100 participants that uses a GOS-E form structure

Here's what I found:

- [11 Studies \(2 top matches\)](#)
- [6 Forms \(1 top match\)](#)

Learn more about these results

Select the relevant forms and studies and click Select Data.

Did we find what you were looking for? ☒ Yes ☐ No

Search with AI

Clear Search

Search with AI using Natural Language

What are you looking for?

Describe what you're looking for within form structures, studies, and data elements, and provide details around specific data you'd like to see. Please check the results for accuracy.

A study with at least 100 participants that uses a GOS-E form structure

Here's what I found:

- [11 Studies \(2 top matches\)](#)
- [6 Forms \(1 top match\)](#)

Learn more about these results

Based on your query for data on anxiety in elderly subjects, here is a comprehensive response:

Primary Form Structure:

- Form: Generalized Anxiety Disorder (GAD-7)
- Form ID: 296
- Justification: This form is specifically designed to screen and measure anxiety symptoms, making it ideal for analyzing anxiety in elderly populations.

Primary Data Element:

- Data Element: 'GAD7WryTooMchScale'
- Reason: Directly measures anxiety symptoms using the GAD-7 scale, with a specific focus on worrying - a key indicator of anxiety particularly relevant to elderly subjects.

Filter Criteria:

- Data Element: 'AgeYrs'
- Filter Expression: Subjects aged 65 years and older
- Justification: Allows precise selection of elderly participants by capturing age in full years

Relevant Studies: The following studies were identified as potentially containing relevant anxiety data:

1. FITBIR-STUDY0000259
2. FITBIR-STUDY0000242
3. FITBIR-STUDY0000243
4. FITBIR-STUDY0000244
5. FITBIR-STUDY0000245
6. FITBIR-STUDY0000247
7. FITBIR-STUDY0000250



Specific AI Projects in Development at BRICs: AI for Dataset Queries

See explanation of why a result is a match

Top Matches:

Select All Top Matches

☐ +

AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury

PI: Hilaire J Thompson

20 272

[Why is this a top match?](#)

☐ +

Adding Legacy Clinical Data to the Federal Interagency Traumatic Brain Injury: Citcoline Brain Injury Treatment Trial

PI: Ross Zafonte, DO

56 1213

[Why is this a top match?](#)

Top Matches:

☐ +

AIM: TBI Impact of Aging on the Immune Response to Traumatic Brain Injury

PI: Hilaire J Thompson

20 272

[Why is this a top match?](#)

☐ +

Adding Legacy Clinical Data to the Federal Interagency Traumatic Brain Injury: Citcoline Brain Injury Treatment Trial

PI: Ross Zafonte, DO

56 1213

[Why is this a top match?](#)

Other Results:

☐ +

B Subphenotypes

PI: Richard Rubenstein

46 346

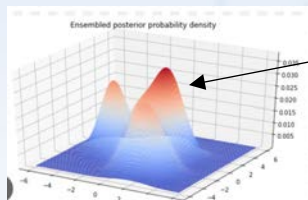
[Why is this a match?](#)

Why this is a Top Match

This is another strong candidate from the GAD-7 scale, measuring difficulty relaxing, which is a key anxiety symptom in elderly individuals

What can be done moving forward: Technical Implementation

- **Clustering of unharmonized data element dictionaries** from current repositories (Ex: Open Data Commons for Spinal Cord Injury) using text embedding models



High Frequency
Concept Cluster – Which variables can
be collapsed to a single
Common Data Element

- **Implementation of Interactive, Collaborative Platforms for CDE curation** that utilize the automated organization afforded by AI Models.
- **Create an AI interactive personal assistant to help researchers explore the repository and plan for their next study:**
 - Suggesting common data elements/forms for their research goals
 - Recommending other studies to consider using those data elements.
 - Automatically aggregating and suggesting potential datasets of use.
 - Providing dataset visualizations and flagging potential sampling bias of a given dataset.
- **Consider federated systems** for data storage
- **Design new tools, workflows, and standards for archiving image feature extraction and predictive machine learning pipelines:**
 - ML Experiment Tracking (mlflow, etc)
 - Apps for Easy Archival and Reproduction of Processing Pipelines: Single record containing the entire data workflow (from raw data to published figures) <https://brainlife.io/about/> Pestilli Lab

What can be done moving forward: Governance and Culture

- **Organize a structured process for additional cross-talk and coordination among** the Common Data Element/Repository Leadership Groups across NIH to promote efficiency and eliminate unintentional redundancy in efforts.
- **Encourage the practice of good software development skills in the context of a medical regulatory landscape earlier in the research cycle.**
- **Re-examine traditional funding paradigms** when building long-term, collaborative infrastructure.
- **Create a system that rewards scientists professionally for collaboration, data sharing, and adherence to high-quality data standards.** Encourage respect for team science.
- **Implement more formalized, well-documented, iterative ‘market’ research** when developing data repositories and tools.
- **Design and automatically monitor new key performance indicators** (beyond # of papers) to obtain more quantitative statistics for collaboration and technology transfer.



BRICS

Biomedical Research
Informatics Computing System

True efficiency arises through collaboration, coordination, recognizing and celebrating human genetic heterogeneity, and collectively building effective infrastructure that provides us all with the bandwidth for enhanced creativity.



National
Institutes
of Health



Center for
Information
Technology



Learn more about BRICS at: <https://brics.cit.nih.gov>