

# GenCDE + DIVER

"AI Accelerated Data Harmonization and Discovery"

Platforms

· · · · · · · · · · · · · ·



**O1. Context aware Al** Tested on sparse health data, applicable to other domains.

**O2. Efficient** Orders of magnitude faster and cheaper than competing NIH initiatives.

**O3.** Interoperability Work across silos without sharing protected data.

**O4.** Harmonization Fuzzy matching, AI and connections to databases through a conversational AI. Automating the tedious parts of data science.



•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•
•</t



## From discovery to generative CDEs, how DIVER evolved.

Al-approach to context specific data harmonization across silos from content types to units of measure and acceptable values.

- Automated CDE discovery by fuzzy matching
- Poorly matched abbreviations or text become novel data dictionary entries using biomedical LLM assistance.

This allows us to build a global data dictionary that can be used for rapid dataset harmonization past what others do with grueling matching and tedious effort.

This research can compliment or accelerate other efforts across NIH and beyond.





### Sparse data to harmonized data made easy.



Tabular data in various formats from disparate data silos. These can range from dataset headers, partial or sparse data dictionaries and even abbreviated feature names.



The Al ingests data and builds CDE parameters for each input. Here in a context specific manner multiple agents will identify likely: variable name(s), equivalent abbreviation(s), data types, expected ranges and units of continuous, and expected content types of categorical or ordinal variables populating a database ...



Deployable python scripts evaluate and flag features in each dataset within the silo (automated). Custom python scripts are deployed at each silo to identify and fuzzy match the CDEs with data features per dataset. Features out of tolerance are flagged for exclusion and modification. At each silo, a harmonized dataset is built. This process step may be iterative.



Quality reports are generated per dataset, per variable and across datasets.

Based on the total number of CDEs across datasets of interest, the missingess rates per CDE and across CDEs per dataset are compared, this is an index of data quality without sharing protected data.





Learning system as a database. A CDE database is populated for the customer allowing them to add or modify with minimal effort, facilitating a system that learns easily as new data is imported ..



Federated analyses can now begin. With a report of overlap between datasets for features within tolerances that are also high quality, federated analyses can begin efficiently.



#### Where it started ...

#### You

In a dataset focusing on neurodegeneration and aging in the context of biomedical research, I have a variable abbreviated as "2minwt". Could you please return what this abbreviation likely means, potential aliases for this variable as well as normal values and units of measure. If you could return this in a tabular format like an expert epidemiologist, that would be much appreciated. Thanks!

#### 🜀 ChatGPT

The abbreviation "2minwt" in the context of a dataset focusing on neurodegeneration and aging in biomedical research is likely to represent "2-Minute Walk Test". This test is often used to assess functional mobility, endurance, and physical fitness in various populations, including those affected by aging or neurodegenerative conditions. Here's how the information might be structured in a tabular format:

Abbreviation	Full Name	Aliases	Normal Values	Units of Measure
2minwt	2-Minute Walk Test	2-Minute Walking Test, 2MWT	Varies by age, sex, and health	Meters (m)

#### Explanation of the Table:

- 1. Abbreviation: "2minwt" the shorthand used in your dataset.
- 2. Full Name: "2-Minute Walk Test" the complete name of the test.
- 3. Aliases: These include "2-Minute Walking Test" and "2MWT", which are other common names for the same test.
- 4. Normal Values: The distance covered in the 2-Minute Walk Test can vary significantly based on age, gender, and overall health condition. For instance, older adults or individuals with neurodegenerative diseases might walk shorter distances compared to younger, healthy individuals. Therefore, it's important to refer to age and health-specific norms.
- Units of Measure: The performance in the 2-Minute Walk Test is typically measured in meters, indicating the distance covered in the given time.

Note that for specific normal values, you may need to consult clinical guidelines or research that provides age, sex, and condition-specific normative data for this test. The values can vary widely depending on the population being studied.

#### Where it went ...

# Prompt resting
Tor Intex, Fow In ConducterFunds(), Cocal-Relicity, Desce Froncessing Fons ).
# Enhanced 'title' prompt handling with additional context
<pre>if pd.isna(df.loc[index, 'title']) or df.loc[index, 'title'] == '':</pre>
title prompt = (
"Your task is to output an acceptable data element title for the variable below using the provided additional context: \n"
f"variable name: ### {df.at[index, 'variable name']} ###\n"
# Adding additional context if available
additional_context =
at organized an oricolous and period (index, organized )):
different descriptions to define and interact and anticidity of anticidity of the description (1).
additional context += f"short Description: {df_at[index, 'short description']}}n"
if 'preferred question text' in df.columns and pd.notna(df.at/index. 'preferred question text');
additional context ++ f"Preferred Question Text: {df_at{index, 'preferred question text'}}\n"
<pre>title_prompt += "Additional context:\n" + additional_context</pre>
# Call to GPT-4 Turbo API
<pre>logging.into("prompt for title: (title_prompt)")</pre>
or.lo(index, title )= call_gota_chat(title_prompt, api_key)
# Handling short description prompt and adding available context
if pd.isna(df.loc[index, 'gpt4 short description']) or df.loc[index, 'gpt4 short description'] '':
short_desc_prompt = (
f"Instruction: Respond with a short description (5-25 words) of the delimiter marked biomedical term below within the clinical context of (subject) that enhances understanding. Additional context is included below the term of interest.\n"
f"Biomedical term: ### (df.at[index, 'title']) ### \n"
# ADDINg UDICIDAL CONCOL 1 AVELADIC
and a contract of a stand of a st
and the second s
if 'short description in of columns and part description']):
additional context ++ f"Short Description: {df.at[index, 'short description']}/n"
if 'preferred question text' in df.columns and pd.notna(df.at[index, 'preferred question text']):
additional_context ++ f"Preferred Question Text: {df_at[index, 'preferred question text']}\n"
if additional_context:
<pre>short_desc_prompt += "Additional_context:\n" + additional_context</pre>
Annual to Dell'Annual for takent for takent for a second 33
Adjungtanovi prosperior and eleverative of the elev
# Handling of keywords prompt and adding available context
if pd.isna(df.loc[index, 'gpt4_keywords']) or df.loc[index, 'gpt4_keywords'] -= '':
kaywords_prompt = (
f"Instruction: Provide a comma separated list of keywords and keyterms for the delimiter marked biomedical term below to optimize its fitness for searching and matching of similar or related elements within the clinical context of [subject].
"the only permissible response format is the following: keyterm, keyword, etc \n"
<pre>f*Biomedical term: ### (df.at[index, 'title']) ### \n"</pre>
* ADLING BUDILICUTIL COTICAL LT BAVALADIO
if 'definition' in draphase and notaclife at index. 'definition' by
additional context as findinition: (if at index 'idditition')))o



#### Where it is ...

Submitted DD	Total N Elements	N GPT4 Processed	Current Hit Rate (%)
BTRIS(Biomedical Translational Research Information System)	53,631	29942	100
CTDB (Clincial Trial Database)	1,943	1,840	77.7
Exercise Log	61	61	100
Intake Form 20n0153	283	282	Auditing
UPDRS (Unified Parkinson's Disease Rating Scale)	285	125	89.9
Metabolic	41	38	84.5
Resting State	101	97	60
Eligibility	14	11	100
Geriatric Depression Scale	20	16	100
Gastrointestinal Diary	12	8	72.5
GSRS (Gastrointestinal Symptom Rating Scale)	19	15	66.6
Tug	8	8	96.3
Heart Rate Variability Daily	18	14	50
Ninehole	11	7	94.3
In-patient Exercise	10	6	100
Kubios Heart Rate Variability	31	27	72.3
MOCA (Montreal Cognitive Assessment)	18	16	93.3
CARD	627	621	82.3
MNPQ	18	14	69.2
NDSR (Nutrition Data System for Research)	124	120	85
NMSS (Non-Motor Symptoms Scale)	103	100	72.7
Pronutra	16	12	100
Psytoolkit	54	36	100
SymPsychToolKit	41	26	Auditing
Safety Monitor Survey	27	24	100
Screening Survey	16	13	100
RADC	79	79	90
Health Rhythms & Inferences	88	88	Auditing
Neuropsych	1,397	1,396	77.6
NLM CDE Repository	23,529	5,961	82.5
TOTAL	82,625	41,003	85.80

# Dozens of sparse

### datasets tested.

- Hot off the press test!
- Generally greater than 85% hit rate (where no edits were needed to any part of the CDE inventory).
- Accomplished months of analyst work in a single day.
- Highly parallelized code but still cost efficient.
- PHI secure.

DATATECNICA

#### What's going on ...

variable_name	gpt4_title	gpt4_unit_of_measure	gpt4_aliases	value_format	standardized_value
			Adult Heart Checkup - Drawing - Max Weight, Adult Cardio Health Follow-		
			up - Blood Test - Peak Weight, Adult Heart Health Monitoring - Drawing -		
			Topmost Weight, Cardiovascular Checkup Adult - Lab Test - Heaviest		
AHHF DRAWING HEAVIE			Weight, Heart Health Evaluation - Blood Work - Maximal Weight, Adult		
ST_WT		kilograms	Cardiac Follow-up - Phlebotomy - Highest Weight, Cardio Exam Adult -	continuous	[1,635]
			Adolescent Drawn Maternal Cardiac Health, Teen Illustration of Mother's		
			Heart Health, Adolescent's Mother's Heart Health Drawing, Teenage		
			Picture of Mom's Heart Health, Adolescent Depiction of Maternal Heart		
			State, Teen's Maternal Heart Health Art, Adolescent's Mother's Cardio		
AHHF_DRAWING_MOM		N/A	Health Drawing	categorical	["1", "2", "3", "4", "5", "6", "7", "8", "Adopted", "Not Done"]
			Adolescent Heart Health Father Drawing Test, AHHFD Assessment,		
			Adolescent's Perception of Father's Heart Health Drawing Assessment,		
			AHH Father Articulation Assessment, Adolescent Constructed Image of		
AHHF_DRAWING_FATHER		N/A	Father's Heart Health Assessment	categorical	["1", "2", "3", "4", "5", "6", "7", "8", "Adopted", "Not Done"]
			considered to consider the state of the stat		
			Acceptable Recording Duration (Days), Reasonable Data Recording Time		
REASONABLE_RECORD_	Reasonable Recording Duration in		(Days), Biomedical Data Record Time (Days), Appropriate Record Duration		
NUMBER_DAYS	Number of Days	Days	(Days), Permissible Data Recording Period (Days)	continuous	[1, 365]
			Day 2 Satiation Time, Satiation Duration Day 2, Day 2 Fullness		
	Time to Achieve Satiation on Day 2 of		Achievement Time, Second Day Satiation Time, Time to Fullness Day 2, Day		
DAY_2_SATIATION_TIME	Satiation Study	minutes	2 Satiation Study Duration	continuous	[0, 720]
			"Devid Certists Operational" Find of Test 41" "D4 Find one Otest Test 4 Find"		
			Day I Satiety Onset until End of Test 1, DI Fullness Start - Test 1 End ,		
			Tist Day Feeling Full until End of 1st Test , Start of Satiety - End of First		
DAY & CATIETY TIME &	"Day 1 - Onset of Satiety - End of Test	Time	lest Day 1, "Day 1: From Fullness Commencement to First Test End",		[00-01 00-50]
DAY_1_SATIETY_TIME_1	Time 1	Time	Onset of Satiation to End of Testing Day 1	continuous	[00:01, 23:59]
			Day 1 Pro Activity 1 Stomach Noison, D1 Pro Act 1 Stomach Disturbances		
			1et Day Pro Event Stomach Pumbling Activity 1 Day 1 Pro Pumbling Day 1		
DAY 1 STOMACH PUMPI	Day 1 Stomach Pumbling Prior to		Pro Activity Costric Noises, D1 Stomach Pumble Before Act 1, Day 1		
E PRE 1	Activity 1	N/A	Borborygmus Pre-Activity 1	hinany	[Ves No]
C_FRC_1	Activity 1		Pre-Procedure Anxiety Level Day 1 Day One Pre-Procedure Anxiety Degree	Uniary	
			Pre-On Anviety Level Day One Anviety Score Pre-Procedure Day 1 Pre-		
			Procedure Anviety on Day One, Anviety Degree Pre 1, Anviety Level Day 1 Pre-		
DAY 1 DEG ANXIETY PR	Degree of Anviety Pre-Procedure on	Unit of Measure' Anviety	Procedure Day 1 Anviety Before Procedure Day One Before Procedure		["No anviety" "Minimal anviety" "Mild anviety" "Moderate anviety" "Severe
E 1	Day One	Rating Scale Score	Anxiety Degree	categorical	anxiety". "Extreme anxiety"]
		in the second second			
			Hunger-Level Post 1 (Day 1), Hunger Degree after 1st Trial (Day 1), Post-		
			Trial 1 Hunger Degree (Day One), Hunger Rating 1st Trial (Day 1), Day 1		
DAY 1 DEGREE HUNGER	Degree of Hunger After First Trial on		Post-Trial Hunger Level, First Day Post-Trial Hunger Degree, Hunger Score		
POST 1	Day One	Numeric Scale (1-10)	after First Trial (Day One), Degree of Hunger - Trial 1 (Day 1)	categorical	["1", "2", "3", "4", "5", "6", "7", "8", "9", "10"]

DATATECNICA

### Interoperability score

Per study per variable interoperability score. This flags variables for site level follow-up by analysts.

The score includes the sum of:

- Missingness rate (%) all
- Threshold tolerance (%) continuous
- Binary coding (%) dummies
- Fuzzy match (%) categorical





## Why does this CDE platform accelerate research?

#### Initial test at NIH has shown that 85% or move of queries need no manual editing.

- Evaluated by multiple NIH collaborators with greater than expected success.
- Poorly matched abbreviations or text become novel data dictionary entries using biomedical LLM assistance.
- Long term of NIH-related research benchmarks accomplished in less than a week of parallel computing and automation by comparison, including a manual audit phase.

#### Exceptional handling of data silos.

- Ingest sparse data dictionary, dataframe headers or just simple abbreviations from analyses.
- Distributed (re-)formatting and outlier flagging.
- Central CDE database for different research areas and contexts is also valuable reference.

# This workflow can be deployed in healthcare and beyond to areas such as finance, climate change, politics etc.

Flexible and modular.

. . . . .

• Employs multiple context aware agents that are modifiable.





### DIVER query tool demo @ CARD, now rolling out!







### DIVER query tool screen capture (to save time)



#### II DIVER GoogleForms Query Tool

① Please Select Login/SignUp From the side bar

General	
ATAC seq Project Name	
Clinical Project Name	
DNA Array Project Name	
Exome Project Name	
Genotyping Project Name	
HiC seq Project Name	
Imaging Desiget Name	
General	~

#### Selected Table: General of Size: (24, 12)

	Timestamp	Study_ID	Study Name	Study Subjects	Study NIH Contact	Study Document Path
Z	9/21/2023 10:10:13	неациявс	Health ABC Study	anımaı	mn@nin.gov	NA
3	9/21/2023 12:38:39	HAAS New	HAAS New Project	cells	abc@abc	NA
4	9/20/2023 15:45:16	NG00017	OHSU GWAS	human	CARD_Diver@list.nih.gov	README: https://drive.google.com/file/d/17vDBKLqwwjBu-C6q04
5	9/21/2023 15:04:57	NN00001	NN ADRD	animal	abc@abc	NA





## **DIVER chat interface (previous beta)**

A Home all CSV Data Ingest La Chat With Diver README
Clear conversation history
B How can I help you?
Ist all target disease areas
The target disease areas are Alzhemier's disease, Dementia general/MCI, Healthy Aging, Vascular dementia/Stroke, COVID-19, Corticobasal syndrome, Frototempoal dementia, Huntington, No disease focus, Parkinson's disease, Gastric Cancer, Diabetes, Cancer, Hepatitis, Uveitis, and HIV.
S List all studies with focus on Alzheimer with metadata
The studies with focus on Alzheimer and their metadata are NG00017, OHSU GWAS, Study Subjects: human, Study NIH Contact: <u>CARD_Diver@list.nih.gov</u> , Study Document Path: README: <u>https://drive.google.com/file/d/17vDBKLqwwjBu-C6q04vUOOZc6jmGD8hQ/view?usp=share_link;</u> Description: <u>https://docs.google.com/document/d/1DOk3NHQtouecrtqmzXkzTWlwRPPBiEp1XJrXeqbAS5g/edit?usp=share_link</u> , Study Description: "Alzheimer's Disease pathology in cognitively healthy elderly: A genome-wide study. OHSU is a GWAS dataset used in the discovery stage of the ADGC Naj et al. meta-analysis from 2011. It is from an cross-sectional and observational study that has 2 study arms and 595 samples. It is one of the ten cohorts assembled by ADGC to identify genetic variants in LOAD.", Study Disease Focus: Alzhemier

What is this data about?







