AI Powered Mapping of Trial Outcomes to CDISC Standards: Unlocking the Potential of Past Data

Meet the Speakers

Ece Kavalci

Title: Machine Learning Engineer

Organization: Lindus Health

Ece is a data scientist with a deep focus on innovative trial design. She holds an MSc in Data Science from King's College London. Adopting a data-centric approach, she has experience on projects such as AI powered study document creation, predictive risk modeling, and advanced analytics for trial design. Her specialized expertise has significantly enhanced trial design and monitoring processes.

Fun Fact: before getting into computer science Ece was an architect working on parametric design and digital fabrication

Oskar Wroz

Title: Software Engineer

Organization: Lindus Health

Oskar is a full stack developer with experience helping build a wide range of health tech products, including a VR simulation platform used by the NHS, AI-enabled drug discovery tools, and, currently, technical solutions that rethink and improve clinical trials. Self-taught as a developer, Oskar holds a Bachelor of Commerce from the University of British Columbia.

Fun fact: before software development Oskar was a musician in a touring indie rock n' roll band.

Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the authors and do not necessarily reflect the official policy or position of CDISC.*

- *The authors have no real or apparent conflicts of interest to report.*

Agenda

1. **Intro:** Current State of Outcome Data
2. **Data Preparation**
3. **Our Approach**: Leveraging LLMs
4. **Demo**: The Solution In Action
5. **Implementation Details**
6. **Takeaways and Limitations**
7. **The Future**

Introduction

What is the state of clinical trial outcome data, and how can AI improve its usefulness?

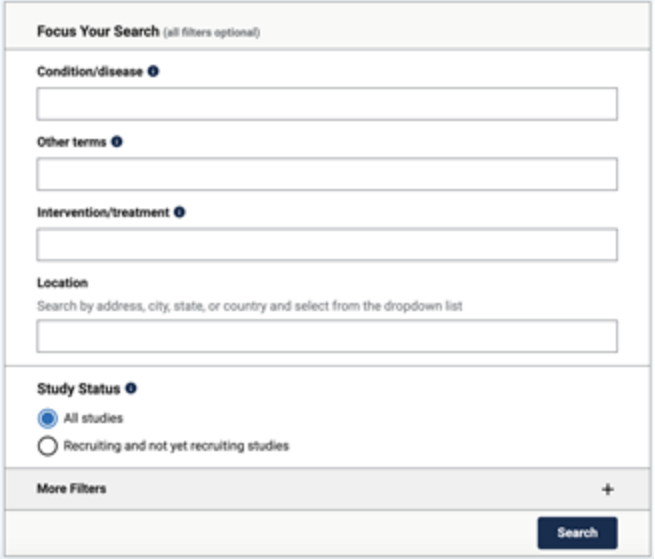
Outcome data is critical but inaccessible

- Clinical trial data is a critical resource for improving research and trial design.
- Much of this data is locked in free-text formats, making it difficult to leverage.

cdisc

Searching for outcomes is a highly manual process

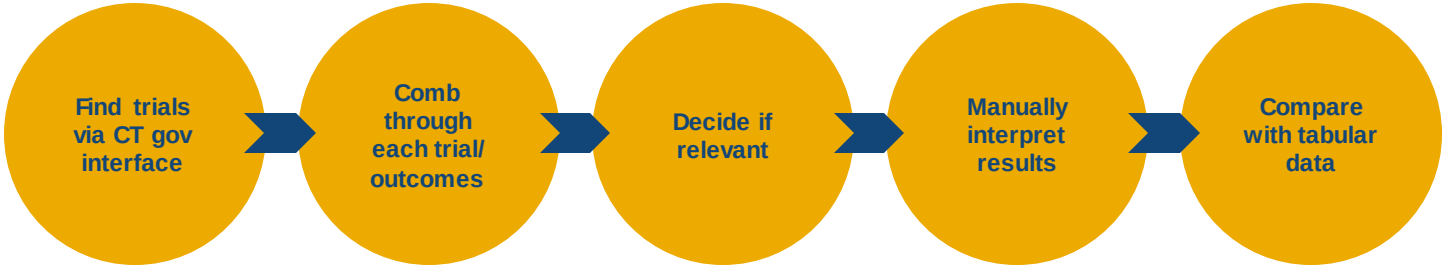
- Searching for outcomes means looking at long lists from individual trials
- Making sense of free text is time consuming



clinicaltrials.gov search interface

cdisc

Searching for outcomes is a highly manual process



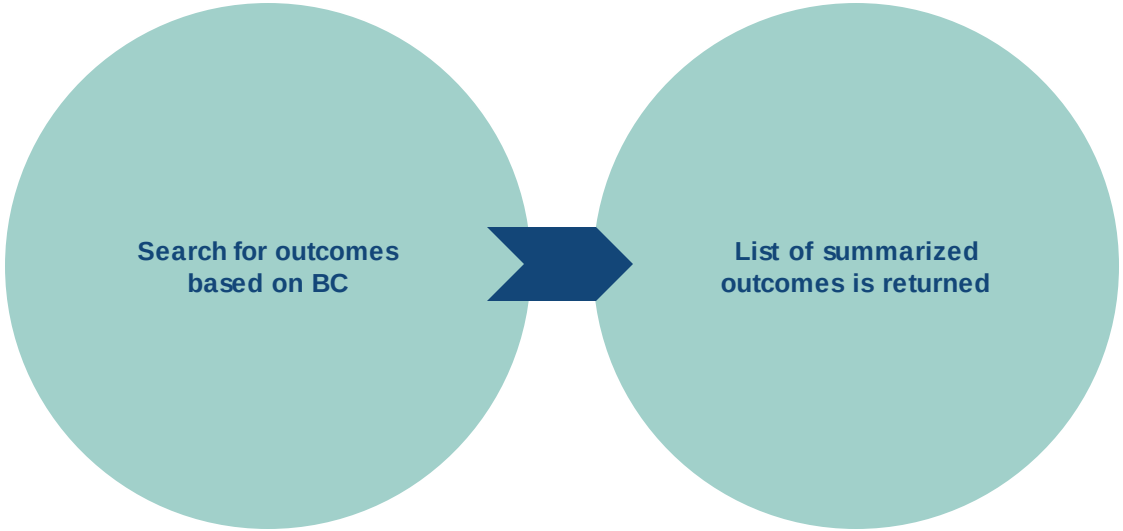
cdisc

AI unlocks outcome data potential

- **AI solution**: Use LLMs to convert free-text outcomes into structured, queryable information.
- **Application**: Map outcomes to standardized definitions (BCs), enabling searchability.

cdisc

AI streamlines the entire process



cdisc

Easy-access data enables better trials

- **Benefits**: accelerates both research and trial design by automating the search and screen process.

- **Ultimate Goal:** take AI assisted trial design even further - generate more intelligent summaries, recommendations, and protocols themselves

cdisc

AI can help build CDISC biomedical concept library

- Attempting to map entities that LLMs identify to Biological Concepts in the CDISC API helps identify potentially missing BCs.

Data Preparation

Preprocessing and joining multiple data sources

AACT Database

Data Filtering:

- By disease area or multiple conditions (i.e. alzheimer's disease)
- Study start year (i.e. all studies started after 2020)

	id	nct_id	outcome_type	title	description	time_frame	population
0	50502000	NCT04096417	Secondary	Overall Survival (OS)	Overall survival (OS) is defined as the time from	29.4 Months	
1	50714722	NCT00472212	Secondary	Change in Accommodative Response From Baseline to Afte	Accommodative response, Right eye. This is a n	6 week outcome exam	
2	50400995	NCT01136785	Secondary	24-hr Profile of Plasma Growth Hormone	The mean plasma growth hormone level will be	after 1 week of active CPAP therapy in the laboratory	
3	50502001	NCT04096417	Secondary	Quality of Life (QOL) as Measured by the LASA [Item 1: Over	Quality of Life (QOL) was measured using item 1	9 Months	All patients th
4	50502002	NCT04096417	Secondary	Incidence of Adverse Events	Adverse events will be summarized by frequenc	5.4 Months	
5	50502003	NCT04096274	Primary	Percentage Fidelity to the OQ-A System Experienced by the	Fidelity to the OQ-A will be measured by using e	0-6 months after youth's baseline/ entry into treatment	Youths from t
6	50502004	NCT04096274	Primary	Change From Baseline to 6-months in Youth Total Problems	The SAC Total Problem Score is a 48-item meas	0-6 months after youth's baseline/ entry into treatment	Analysis inclu
7	50502005	NCT04096274	Primary	Percentage Fidelity to the OQ-A System Experienced by the	Fidelity to the OQ-A will be measured by using e	0-6 months after youth's baseline/ entry into treatment	Analyses bas
8	50502006	NCT04096274	Primary	Change From Baseline to 6-months in Youth Total Problems	The SAC Total Problem Score is a 48-item meas	0-6 months after youth's baseline/ entry into treatment	One clinic wa
9	50502007	NCT04081233	Primary	Hospital Length of Stay	Number of days patient is in the hospital	180 days after admission	
10	50502008	NCT04081233	Secondary	Mortality	Death following trauma injury involving rib fractu	180 days after admission	

CDISC's Biomedical Concepts

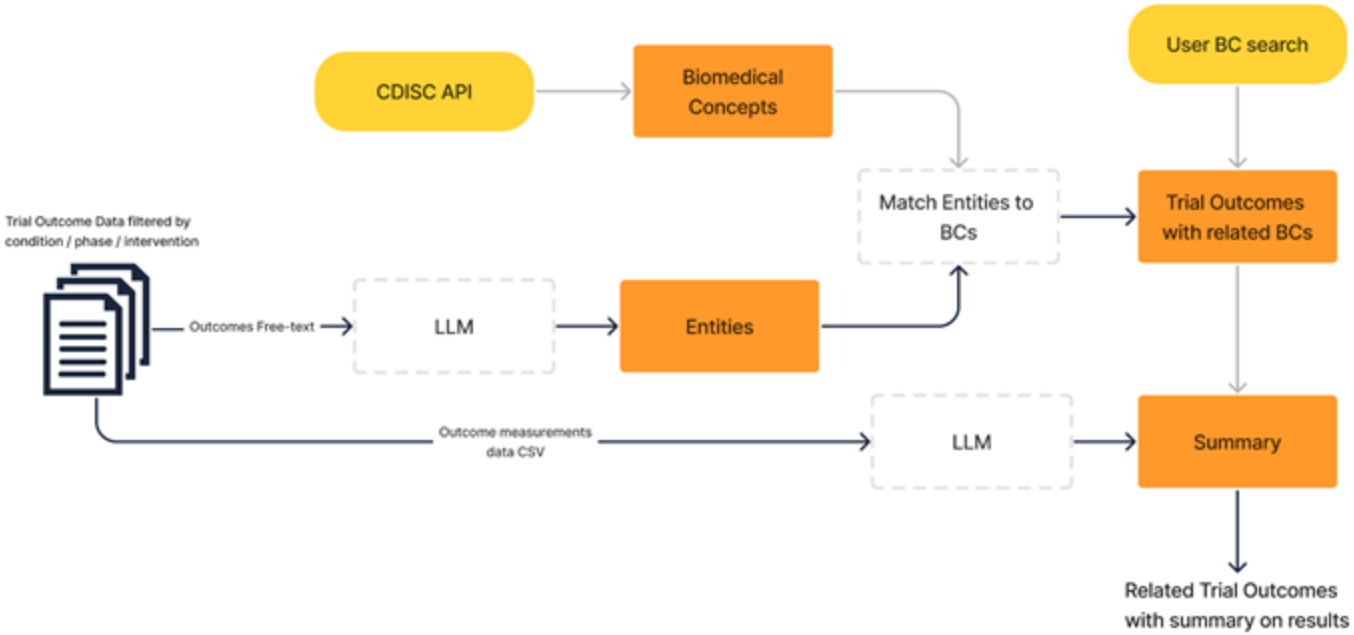
- Based on existing ontologies like NCIt
- CDISC API allows us to access and extract biomedical concepts


```
{
        "conceptId": "C60832",
        "shortName": "Oxygen Saturation
Measurement",
        "definition": "The measurement of the
ratio of oxygenated hemoglobin to total
hemoglobin in the blood.",
        "href":
"https://ncithesaurus.nci.nih.gov/ncitbrowser/Co
nceptReport.jsp?dictionary=NCI_Thesaurus&ns=ncit
&code=C60832",
        "categories": [
            "Vital Signs",
            "Oxygen Saturation Measurements",
            "Oximetry Tests",
        ], …}
```

Our Approach

LLM powered entity-BC mapping and outcome summarising

Approach


CDISC API
Biomedical Concepts
User BC search
Match Entities to BCs
Trial Outcomes with related BCs
Trial Outcome Data filtered by condition / phase / intervention
Outcomes Free-text
LLM
Entities
Outcome measurements data CSV
LLM
Summary
Related Trial Outcomes with summary on results

Demo

An interface for exploring trial outcomes

Ontobee

Home | Intro | Statistics | SPARQL | Ontobee | Annotator | Tutorial | FAQs | References | Links | Contact | Acknowledge | Issues

NCI Thesaurus OBO Edition

Keywords: Search terms

Class: Adverse Event

Term IRI: http://purl.obolibrary.org/obo/NCIT_C41331

Definition: Any unfavorable or unintended disease, sign, or symptom (including an abnormal laboratory finding) that is temporally associated with the use of a medical treatment or procedure, and that may or may not be considered related to the medical treatment or procedure. Such events can be related to the intervention, dose, route of administration, patient, or concurrent or interaction with another drug(s) or procedure(s).

[Definition Source: NCI]

Annotations

- **ALT_DEFINITION:** An unexpected medical problem that happens during treatment with a drug or other therapy. Adverse effects do not have to be caused by the drug or therapy, and they may be mild, moderate, or severe. Any unfavorable medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event (AE) can therefore be any unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product. NOTE: For further information, see the ICH Guideline for Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. [Modified from ICH E2A] Synonyms: side effect, adverse experience. See also serious adverse event, serious adverse experience. CONTE: (paraphrase) Any unfavorable and unintended sign, symptom, disease, or other medical occurrence with a pharmaceutical preparation with this use in a medical product, procedure or other therapy, or in conjunction with a research study, regardless of causal relationship. EXAMPLE(S): death, back pain, headache, pulmonary embolism, heart attack. Any unfavorable medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event (AE) can therefore be any unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product. NOTE: For further information, see the ICH Guideline for Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. [After ICH E2A] See also medical adverse event, serious adverse experience. Any unfavorable and unintended sign, symptom, disease, or other medical occurrence with a temporal association with the use of a medical product, procedure or other therapy, or in conjunction with a research study, regardless of causal relationship. EXAMPLE(S): death, back pain, headache, pulmonary embolism, heart attack. [OTHER SAMPLE(S): NCI1333. The BRIDG 2.02 Rule 4 Ontology Tracker issue [BO1384] (referring a reader to website that supplemented that which Adverse Event did not be specified by a Safety/Surveillance). Any unfavorable or unintended medical sign, or symptom (including an abnormal laboratory finding) that is temporally associated with the use of a medical treatment or procedure, and that may or may not be considered related to the medical treatment or procedure. Such events can be related to the intervention, dose, route of administration, patient, or concurrent or interaction with another drug(s) or procedure(s).
- **Contributing_Source:** CDISC; FDA; NCI; PMA; CDISC; NCI; CTCAE; CTCAE; MedDRA
- **Has_DOC_Value:** Myositis, CTCAE 5.0: Dry Skin, CTCAE; Confusion, CTCAE 5.0: Tumor Pain, CTCAE 5.0: Sinus Pain, CTCAE 5.0: Tremor, CTCAE 5.0: Anxiety, CTCAE 5.0: Sinus Tachycardia, CTCAE 5.0: Joint Effusion, CTCAE 5.0: Ventricular Cardiac Arrhythmia, CTCAE 5.0: Photophobia, CTCAE 5.0: Hypotension Pain, CTCAE 5.0: Fatigue, CTCAE 5.0: Blurred Photophobia, CTCAE 5.0: Eye Pain, CTCAE 5.0: Fallopian Tube Obstruction, CTCAE 5.0: Dysmenorrhea, CTCAE 5.0: Dysgeusia, CTCAE 5.0: Dysphonia, CTCAE 5.0: Cranial Nerve Pain, CTCAE 5.0: Cognitive Disturbance, CTCAE 5.0: Hydrophobia, CTCAE 5.0: Agitation, CTCAE 5.0: Lymph Node Pain, CTCAE 5.0: Stomach Pain, CTCAE 5.0: Erectile Dysfunction, CTCAE; Insomnia, CTCAE 5.0: Scalp Pain, CTCAE 5.0: Urinary Pain, CTCAE 5.0: Photophobia, CTCAE 5.0: Osteoporosis, CTCAE 5.0: Hemoglobinuria, CTCAE 5.0: Back Pain, CTCAE 5.0: Memory Impairment, CTCAE 5.0: Bruises, CTCAE 5.0: Lip Pain, CTCAE 5.0: Angioma, CTCAE; Wolff-Parkinson-White Syndrome, Toxic Epidermal Burn, CTCAE; Urethral Hemorrhage, CTCAE 5.0: Cutaneous Pain, CTCAE 5.0: Ischial Pain, CTCAE 5.0: Testicular Pain, CTCAE 5.0: Pleuritic Pain, CTCAE 5.0: Propelvic Pain, CTCAE 5.0: Uropelvic Pain, CTCAE; Non-Reflexive, CTCAE 5.0: Photosensitive Pustules, CTCAE 5.0: Serum Sickness, CTCAE 5.0: Flushing, CTCAE 5.0: Gastrointestinal Fistula, CTCAE 5.0: Rectal Perforation, CTCAE 5.0: Small Intestinal Perforation, CTCAE 5.0: Gastrointestinal Hemorrhage, CTCAE 5.0: Gum Perforation, CTCAE 5.0: Rectal Pain, CTCAE 5.0: Uncontrolled Pregnancy, Oral Pain, CTCAE 5.0: Depression, CTCAE; Ventricular Arrhythmia, CTCAE 5.0: Hairless Arm Hyperkeratosis, CTCAE; Myalgia, CTCAE; Anorexia, CTCAE; Constipation, CTCAE; Colitis, CTCAE; Urinary Retention, CTCAE; Aspiration, CTCAE; Vaginal Fistula, CTCAE; Weight Loss, CTCAE; Nausea, CTCAE; Weight Gain, CTCAE; Obesity, CTCAE; Hypothermia, CTCAE; Spermatic Cord Hemorrhage, CTCAE; Rectal Hemorrhage, CTCAE; Pharyngeal Hemorrhage, CTCAE; Pancreatic Hemorrhage, CTCAE; Prostatic Hemorrhage, CTCAE; Ovarian Hemorrhage, CTCAE; Oral Hemorrhage, CTCAE; Mediastinal Hemorrhage, CTCAE; Hydrocephalus, CTCAE; Uterine Hemorrhage, CTCAE; Vaginal Hemorrhage, CTCAE; Testicular Hemorrhage, CTCAE; Tracheal Bronchial Hemorrhage, CTCAE; Laryngeal Hemorrhage, CTCAE; Gas Hemorrhage, CTCAE; Vaginal Hemorrhage, CTCAE; Hematoma, CTCAE; Rhinitis, CTCAE; Tear

Implementation Details

Using LLMs for extraction, mapping and summarising tasks

LLMs outperforms NER models for entity extraction

Trial Outcome: “Change From Baseline in Hematology Parameter: Erythrocyte. Mean Corpuscular Hemoglobin (Ery. MCH)”

NER Model Output

```
'entities': [
(Ery, 92, 95, 'DRUG'),
(MCV, 97, 100, 'DRUG')
]}
```

LLM Output

```
"entities": [
        "Change From Baseline",
        "Hematology Parameter",
        "Erythrocyte",
        "Mean Corpuscular Hemoglobin",
        "Ery. MCH",
    ],
```

LLM powered Entity - Biomedical Concept mapping

```
{       "outcome": "Part 2: Change From Baseline in
Chemistry Parameters: … Cholesterol, Creatinine, Direct
Bilirubin, Glucose, HDL Cholesterol, …",
        "entities": [
            ...
            "Cholesterol",
            "Creatinine",
            "Direct Bilirubin",
            "Glucose",
            "HDL Cholesterol",
            ...
        ],
    },
```

```
    "conceptId"  "TBD"
    "shortName"  "Direct Bilirubin"
    "definition"  "The portion of
bilirubin that is directly processed by
the liver ..."
    "href"  ""
    "categories"   "Laboratory Test
Result"  "Liver Function Test"
    "_links"
    "synonyms"   "Conjugated Bilirubin"
    "resultScales"   "Quantitative"
    "coding"
    ...
```

Using LLMs Tabular Data Summary

- For each trial outcome that is returned by the user's biomedical concept search, a summary of outcome measurements is added.
- LLMs are getting better at summarising/analysing tabular data and this was an attempt to showcase how they can automate data analysis.
  - *this is of course far from statistical analysis, and just for summary purposes*.
- A step further could be to automate graph generation instead of text summaries, but this would be computationally expensive in its current form.

Takeaways and Limitations

Defining missing BCs for CDISC library

- In its current state, the BCs are very limited, and therefore don't cover a significant portion of trial outcomes.
- This approach could be developed into defining BCs that are missing from the CDISC library.
- Trial outcome entity - BC mappings unlock a potential to structure any kind of free-text data (i.e. eligibility criteria).

Limitations

- LLM output formats are not 100% reliable, which requires extra checks
- Agents with additional steps could prepare output in an expected way
- Currently using public data - more work required to use private data

The Future

What are the next steps to fully unlock trial outcomes?

Future improvements are inevitable

- As LLMs and specialized models improve, we can rely on them for deeper understanding of trial outcomes, and, likely, to assist the actual design of new trials.

- As standardized library of BCs grows, more can be identified in outcomes

cdisc

Thank You!

SPROUT

Select a Study

Alzheimer's Test Study

Select

Create New Study

Explore

Trial Outcomes

Demo


SPROUT
Back to Studies
Search a Biomedical Concept
Search a biomedical concept here...
Check Outcomes

Demo

Heart Rate

Check Outcomes

Results:

NCT05074498

Trial Outcome

Part 1: Change From Baseline in Heart Rate

Matched Biological Concepts

('Heart Rate', 'C49677')

Matched Entities

Show more

Summary

Show More

NCT05074498

Trial Outcome

Part 2: Change From Baseline in Heart Rate

Demo

Results:

NCT05074498

Trial Outcome

Part 1: Change From Baseline in Heart Rate

Matched Biological Concepts

('Heart Rate', 'C49677')

Matched Entities

Change From Baseline | Heart Rate

Show Less

Summary

Show More

Demo

Results:

NCT05074498

Trial Outcome

Part 1: Change From Baseline in Heart Rate

Matched Biological Concepts

('Heart Rate', 'C49677')

Matched Entities

Show more

Summary

- The clinical trial outcome is identified by the outcome_id 50499227.
- The clinical trial is identified by the nct_id_x NCT05074498.
- The outcome type is "Primary Part 1: Change From Baseline in Heart Rate".
- The outcome measurements are in beats per minute.
- The data includes measurements for different result groups identified by ctgov_group_code OG000, OG001, OG002, and OG003.
- The data includes measurements for different time points within the time frame of Baseline and Up to Day 104.
- The data includes the mean and standard deviation values for each result group and time point.
- The data includes the number of participants (count) for each result group and time point.

- The mean change from baseline in heart rate for result group OG000 is -2.0 beats per minute.
- The mean change from baseline in heart rate for result group OG001 is 3.3 beats per minute.
- The mean change from baseline in heart rate for result group OG002 is 4.0 beats per minute.
- The mean change from baseline in heart rate for result group OG003 is -4.2 beats per minute.
- The standard deviation values for all result groups are within a reasonable range.

Show Less

NCT05074498

Trial Outcome

Appendix 1: Abstract

Shortened Abstract

Clinical trial data is a valuable resource for improving trial design and accelerating research. However, much data remains locked in free-text formats across sources like clinicaltrials.gov, which has outcome data for over 60,000 completed studies. Large language models present an opportunity to unlock this data and transform it into structured, queryable information. This presentation describes an approach that uses AI to map outcome data containing numerical, categorical and free-text columns to standardized endpoint definitions like CDISC Biomedical Concepts. This creates a structured dataset, connects historical data to emerging standards and models, and enables new use cases. Researchers can search outcomes by domain or metric to find precedents to inform trial design. Data can be aggregated for meta-research and benchmarking, and predictive modeling on this harmonized data could optimize future trials. By transforming free-text outcomes into structured endpoints mapped to standards, AI can bring legacy clinical trial data back to life and accelerate research through data-driven trial design.