MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Small Test Set for Dataset Discovery:

Datasets in three diabetes research papers

Version 1

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.

Test set: Curated metadata about datasets in three diabetes papers

This document describes a small, exploratory test set that can be used to assess tools' ability to identify datasets in papers, capture metadata about these datasets, and/or find datasets in papers in response to queries.

This test set consists of metadata and queries about datasets mentioned in three papers about diabetes research. We created this test set by combining LLM-generated output and subject matter expert review, revisions, and curation.

The list of papers and the dataset metadata are in file *MITRE_ASKEM-BDF_Test_set-Datasets_in_diabetes_papers.xlsx.* All three papers are in the PubMed Central Open Access set and use the copyright license CC-BY-4.0.

Criteria for Dataset Inclusion

In identifying datasets for this test set, we considered the following to be within scope: We curated information about datasets in repositories, tables in the bodies of papers, and tables in supplementary materials. We considered images and figures that are not in the form of tables to be out of scope. We also excluded references to datasets from other papers that were not associated with repositories.

For each paper, we have attempted to comprehensively identify every dataset mentioned that is within this scope.

Dataset Details

The first sheet of the excel file, 'papers in set', has PubMed Central identifier, citation information, and a hyperlink to the license for each paper.

The second sheet, 'metadata for datasets in papers', includes the following information for each dataset:

- **PMCID**: PubMed Central identifier for the paper the dataset came from.
- **Dataset Number within Paper**: A number we assigned to each dataset for convenient referencing.
- **Dataset Name:** A name for the dataset, directly taken from the paper if the dataset is named there.
- **Location Type:** Whether the dataset is in a repository, a table in the paper body, or a table in supplementary materials. The repository category is used for everything external to a paper and its supplementary material.
- **Dataset Metadata:** For datasets in repositories, the metadata includes *repository_name*; *dataset_id*, and *url*, to the extent this information is available in the

paper and its supplementary materials. For datasets from tables in the paper body or supplementary materials, the metadata includes *file_name (*e.g., "125_2023_5935_MOESM1_ESM.pdf"), *sheet* (the sheet name for datasets that are in excel files), *dataset_name* (e.g., "ESM Table 1.", and *table_title* (e.g., "Characteristics of control and type-2 diabetic donors included in western blot study") to the extent this information is available in the paper and its supplementary materials .

- **Description:** A one-to-three-sentence description of the dataset that includes the kinds of information the dataset contains, some information about the methods used to generate it, and sometimes the number of samples or cases involved or their characteristics. In the descriptions we avoided paper-specific acronyms that are defined in the paper, and either spelled the words out or used synonyms. Note that the Verbatim Evidence (see below) is not always sufficient to write the descriptions and queries; in particular, we have not included evidence sentences with some contextual information that applies paper-wide, such as the meaning of acronyms that are spelled out elsewhere in the paper.
- **Verbatim Evidence**: Brief, verbatim excerpts from the paper that provide support for the information captured about the dataset. Excerpts may include table titles, single sentences, and excerpts up to a few sentences long. Each dataset has 1-6 pieces of evidence. This evidence is not intended to be comprehensive; for some datasets there is additional supporting evidence within the papers.
- **Query:** A query statement beginning with "Find" that requests the type(s) of data and at least one salient feature about the dataset. An example is "Find data comparing body weight, glucose levels, and insulin levels in type 2 diabetes mouse models and control mice." Queries were written at a level of generality intended to represent realistic searches a scientist might want to do. Consequently, some of the queries are identical or similar to queries for other datasets in this test set, and some of the queries in the set have multiple correct matching datasets.
- **Comments about the dataset:** Notes about some datasets that have less common location features.

Methods Used to Generate this Test Set

We generated this test set by combining LLM extraction of information and human review, revision, and curation. We used GPT-4o to first extract information using the prompt in **Appendix A: Prompt Given to LLM to Identify Datasets in a Paper and write Corresponding Queries**. In addition, we later repeated this with slightly modified versions of this prompt that utilized supplementary materials pdfs in addition to paper body, asked for results in table format, and provided a couple examples. See the sheets with names that start with 'GPT-4o output' for these prompts and the LLM-generated

output from them. For two papers, the paper body text and supplemental file were uploaded together as context; for PMC8991226, these documents were run separately because combining them resulted in missing many datasets.

We skimmed the papers manually, searching for and adding datasets missed by the LLM. We combined the LLM and manually found datasets, and reviewed and edited the LLM-generated output, including revising descriptions, adding and deleting supporting evidence text, and modifying queries. Common revisions included spelling out acronyms, adding critical contextual information, removing tangentially relevant evidence sentences, revising language to be more characteristic of the way researchers write, and making queries more general. For nearly all the datasets returned by the LLM, we modified the description, evidence, and/or query in some way.

Some Potential Ways to Use this Test Set

- Assess a tool's ability to find all datasets in a paper: Compare datasets identified in the papers by your tool with the list here.
- Assess a tool's ability to retrieve metadata about datasets from a paper: Compare dataset metadata (including urls, ids, and descriptions) extracted by your tool to the information listed here.
- Assess a tool's ability to find datasets matching a query: For each query, assess the ability of your tool to return the matching dataset given 1) the one paper the question is from or 2) a larger set of papers from the PubMed Central OpenAccess set. Note that some of the queries have multiple correct matching datasets in this test set. Consequently, a tool's answers (or top answers if ranking is used) should include the dataset paired with the query in the excel file.
- Use this data to develop an LLM-judge prompt to evaluate results at a larger scale (see next section).

Recommendations for Future Test Set Development and Evaluation

Even though we leveraged an LLM to generate first pass results for this test set, the development of this test set still took considerable time and manual effort to complete; this approach would not scale well for creating a large test set. For a larger scale evaluation, rather than creating additional test data we recommend investigating whether an LLM-judge prompt can assist in evaluating tool output. This test dataset could be used as positive examples for evaluation of the LLM-judge. If interested in creating additional test data, we recommend further reducing the manual load by 1) revising the prompt to identify datasets to yield better results, including minor changes like instructing the LLM to spell out acronyms in the description and 2) trying a newer, better LLM to see if it yields suitable test data. We also recommend considering a more select scope for the datasets (e.g., patient or sample metadata, data about molecular results for samples, or descriptive or statistical analysis results).

Appendix A: Prompt Given to LLM to Identify Datasets in a Paper and write Corresponding Queries

MITRE provided the following prompt to Chat GPT-4o to identify datasets in a paper, extract descriptive information, and write corresponding queries.

Use the uploaded paper. Your task is to find and generate a list of ALL the datasets mentioned in the paper and generate a query statement for each dataset. Be sure to include EVERY table in supplementary materials and ALL datasets that the text says are in repositories. Use only information you find in the paper and only list datasets for which there is evidence in the paper.

For each dataset, provide the following information:

- *If it is in a repository, the body of the paper such as in a table, or a supplementary file. If in a repository, provide the name of the repository, a dataset id, and/or a url. If in the paper body or supplementary information, provide a name for the table.*
- *Provide a 1-3 sentence description of each dataset, including the kinds of variables it contains, the methods used to generate it, and the number of samples or cases it includes, if that information is available in the paper.*
- *Provide 2-3 evidence sentences, verbatim from the paper, that provide support for the information you just provided about the dataset. If the table has a descriptive title and/or name, you can list that verbatim as one of the pieces of evidence.*
- *Include a query statement with, at minimum, the type(s) of data, type(s) of tissue; you may include 1-2 other salient features that are notable about the dataset in the query. Start every query statement with "Find". Here are examples of good queries:*
    - *“Find biological sample data linked to electronic health records of patients with diabetes.”*
    - *“Find genomic data associated with type 1 or type 2 diabetes.”*
    - *“Find data on genetic variants linked to type 2 diabetes.”*