Prompt for an LLM tool to Harmonize GEO datasets

Create a harmonized dataset including information about pheochromocytoma and paraganglioma tumor samples from which there is RNA-seq or RNA microarray data.

Use the following GEO datasets (csv files provided):

GSE2841, GSE19422, GSE19987, GSE39716, GSE50442, GSE51081, and GSE67066

Include the following information in the harmonized dataset for each sample, where information is available:

- Accession number for the RNA-sequencing data for each sample
- Patient case identifier
- Source of the sample data, e.g., repository dataset id (e.g., GSE2841)
- Type of cancer (pheochromocytoma or paraganglioma)
- Location of tumor, either adrenal; extra-adrenal; head and neck; or metastases
- Whether the tumor is malignant or benign
- A driver gene with a mutation, in HGNC format. If driver genes were assessed but there were no mutations in them, report “Unknown”
- Patient age
- Patient sex

Create appropriate variables, variable names, and standardized values where appropriate. If data is not available, use “NA”. Include a data dictionary with variable names and descriptions, including units and the standardized values where applicable. Also include a list of the assumptions made in harmonizing the data, e.g. that an abbreviation corresponds to a certain type of cancer.

##

Extensions, with additional challenges:

- Start with the paper Zethoven et al. 2022. Find and download all the GEO files with metadata about the PCPG bulk-RNA-seq data used in the paper
- Integrate and harmonize data from the TCGA-PCPG study with that in the above prompt
- Integrate and harmonize data from the metadata about the single cell RNA-seq data original to Zethoven, in table S1