Description of the Validation Methodology

The **TICTAC-MedlineGenomics Validation** process compares disease-gene associations from **TICTAC** with those from **MedlineGenomics** (formerly Genetics Home Reference) to validate the accuracy and overlap of associations for the same conditions. This involves mapping diseases through **Disease Ontology IDs (DOIDs)** and **UMLS Concept Unique Identifiers (CUIs)**. The process integrates, transforms, and analyzes multiple datasets to evaluate agreement between the two sources.

Steps in the Validation Process

1. Read and Process DO2UMLS Mapping File

- **Input**: DO2UMLS mapping file from Disease Ontology.
  - This file maps DOIDs to UMLS CUIs.
- **Steps**:
  - Load the mapping file.
  - Split `umls_cui` values delimited by vertical bars (`|`) into individual rows for one-to-many relationships.
  - Extract unique counts of DOIDs and CUIs to ensure data integrity.
- **Purpose**: Establish a robust mapping between DOIDs and CUIs to connect disease terminologies used by TICTAC and MedlineGenomics.

2. Read and Process MedlineGenomics Conditions

- **Input**: Medline file containing conditions and their UMLS CUIs.
- **Steps**:
  1. Load the file and split CUIs delimited by commas into individual rows.
  2. Format UMLS CUIs to match the standard format (e.g., prefixing with `UMLS_CUI:`).
  3. Filter out rows without CUIs and extract unique counts of Medline conditions and CUIs.
- **Purpose**: Prepare MedlineGenomics conditions for integration with DO2UMLS mappings.

3. Join MedlineGenomics Data with DOID-CUI Mappings

- **Steps**:
  1. Perform a left join between MedlineGenomics data and DOID-CUI mappings on the `umls_cui` column.
  2. Validate the merged dataset with sample records.
- **Purpose**: Enable comparison of disease-gene associations by aligning Medline conditions with DOIDs via CUIs.

4. Process TICTAC Disease-Gene Associations

- **Input**: TICTAC disease-gene associations from `disease_target_association_with_doid.parquet`.
- **Steps**:

1. Read and format TICTAC data.
    - Ensure the `doid` column is in the correct format (e.g., `DOID:6713`).
2. Select and rename relevant columns (`doid`, `gene_symbol`, `mean_rank_score`) for compatibility.
3. Count unique DOIDs and gene symbols to verify dataset integrity.
4. Join TICTAC associations with DOID-CUI mappings to associate UMLS CUIs with TICTAC data.

- **Purpose**: Prepare TICTAC disease-gene associations for validation against MedlineGenomics data.

5. Read MedlineGenomics Disease-Gene Associations

- **Input**: Medline disease-gene associations file.
- **Steps**:
    1. Load the file and split CUIs into individual rows.
    2. Format CUIs to standard format (`UMLS_CUI:`).
    3. Filter rows without CUIs and validate with sample records.
- **Purpose**: Prepare MedlineGenomics disease-gene associations for comparison with TICTAC data.

6. Compare Disease-Gene Associations

- **Input**: TICTAC and MedlineGenomics disease-gene associations, mapped via UMLS CUIs.
- **Steps**:
    1. Identify common CUIs between TICTAC and MedlineGenomics datasets.
    2. Filter and extract associations for the common CUIs in each dataset.
    3. Count unique CUIs, gene symbols, and associations in the filtered validation sets.
- **Purpose**: Focus the validation on conditions present in both datasets.

7. Measure Agreement

- **Steps**:
    1. Concatenate TICTAC and MedlineGenomics associations for the common CUIs.
    2. Identify duplicate entries to quantify overlaps.
    3. Calculate the percentage of MedlineGenomics associations that are found in TICTAC.
- **Purpose**: Evaluate the extent of agreement between the two datasets.

Key Validation Metrics

1. **Dataset Integrity**:
    - Unique counts of DOIDs, CUIs, gene symbols, and associations are tracked at each step to ensure the correctness of data transformations.
2. **Overlap in Disease Conditions**:

    - Number and percentage of common CUIs between TICTAC and MedlineGenomics are computed.
3. **Overlap in Disease-Gene Associations**:
    - Fraction of MedlineGenomics disease-gene associations found in TICTAC, measured as the percentage of duplicates in the combined dataset.

Summary of Results

- **TICTAC Dataset**:
    - 2243 unique DOIDs, 2022 unique gene symbols.
- **MedlineGenomics Dataset**:
    - 1216 conditions with CUIs and 2142 unique CUIs.
- **Common CUIs**:
    - 193 CUIs shared between TICTAC and MedlineGenomics.
- **Validation Sets**:
    - **TICTAC**: 63569 associations for 193 CUIs, 1804 gene symbols.
    - **MedlineGenomics**: 1247 associations for 193 CUIs, 967 gene symbols.
- **Agreement**:
    - 136 overlapping associations (10.91% of MedlineGenomics associations).

Conclusion

This methodology systematically validates TICTAC disease-gene associations against MedlineGenomics data by leveraging standardized disease terminologies (DOIDs and CUIs). The validation highlights areas of overlap and divergence, providing a measure of consistency between the datasets.