Molecular Targets Platform (MTP)
Data Validation 3 (DV3)

2022-10-04 ZD

Purpose:

This validation will test "completeness and accuracy of data loaded into the platform" and follow MTP Data Validations 1&2. It will compare the data displayed within the platform GUI (after loading) to the expected values within the data (before loading). Automated scripts will pull test cases from the data that can be fed into platform testing automations to check displays for completeness and accuracy.

Scope:

DV3 will focus on displays within the platform that relate to new pediatric data, including those that happen to incorporate Open Targets (OT) data. New pediatric data includes the Food and Drug Administration's Pediatric Molecular Target Lists (FDA PMTL); the much larger collection of evidence data provided by the Children's Hospital of Philadelphia (CHoP); and derived summary tables, such as the Pediatric Cancer Data Navigation (PCDN) page. It will not validate or test displays that only include OT data without pediatric data.

The testing within DV3 will use sampling (spot-checking) methods, though scalability to meet automation capacity will be a design goal. Samples tested will include a set of defined high-profile genes and diseases that we expect will garner an abundance of user attention. We will also include a random sampling of genes and diseases (associated with CHoP data) to expand testing scope.

Input File Descriptions:

1. **CHoP Somatic Alterations files**
    a. Processing levels:
        i. Raw: JSONL files received directly from CHoP
        ii. Processed: JSONL files after any FNL pre-processing (scoring, temporary fieldname fixes, etc.)
        iii. Errored: JSONL files that exclusively contain records that did not successfully load during the MTP ETL process
        iv. Succeeded: JSONL files that exclusively contained records that did not error while loading during the MTP ETL process
    b. Filenames:
        i. CNV (by Gene): gene-level-cnv-consensus-annotated-mut-freq.jsonl
        ii. Fusion by Gene: putative-oncogene-fused-gene-freq.jsonl
        iii. SNV by Gene: gene-level-snv-consensus-annotated-mut-freq.jsonl
        iv. Fusion (by Fusion): putative-oncogene-fusion-freq.jsonl
        v. SNV by Variant: variant-level-snv-consensus-annotated-mut-freq.jsonl

2. **CHoP Gene Expression files**
    a. Raw JSONL files received directly from CHoP for QA, but pulled into MTP via API
    b. Gene-wise Gene Expression tables (evidence page):
        long_n_tpm_mean_sd_quantile_gene_wise_zscore.jsonl
    c. Group-wise Gene Expression tables (target page):
        long_n_tpm_mean_sd_quantile_group_wise_zscore.jsonl

3. **Open Targets Reference files**
    a. OT public FTP path: ftp.ebi.ac.uk/pub/databases/opentargets/platform/{version}/output/etl/json/

    b. associationByOverallDirect/: Direct target-disease associations used in the Target associations page
    c. associationByOverallIndirect/: Indirect target-disease associations used in the Disease associations page
4. **FDA PMTL file**
    a. pmtl_{version}.csv: FNL-curated csv ingested into MTP that represents the PMTL

5. **PCDN file**
    a. chopDataNavigationTable_{version}.json: Derived table containing all unique target-disease combinations within CHoP data, with T/F indicators of datatsource presence

Test Case Generation Overview:

1. Generate the Succeeded CHoP SA files by filtering the Processor (or Raw) files with the Errored files
2. Use the Succeeded CHoP SA files to generate test cases for OpenPedCan SA tabs on both Target and Evidence Pages
3. Use OT Indirect Association file to generate test cases Disease Page associations. This step will need to be redesigned when we incorporate non-zero evidence scoring into CHoP data.
4. Use OT Direct Association file to generate test cases Target Page associations. This step will need to be redesigned when we incorporate non-zero evidence scoring into CHoP data.
5. Use Gene Expression (gene-wise) file to generate test cases for OpenPedCan Gene Expression Evidence page.
6. Use Gene Expression (group-wise) file to generate test cases for OpenPedCan Gene Expression Target page.
7. Use PMTL file to generate test cases for FDA PMTL annotation (Target page) and FDA PMTL page
8. Use PCDN file to generate test cases for PCDN page

Test Descriptions:

1. Target Page
    a. Associated diseases (direct association): Test row count
    b. Target Profile
        i. FDA PMTL annotation: Test value
        ii. OpenPedCan Somatic Alterations
            1. SNV By Gene tab: Test row counts
            2. SNV By Variant tab: Test row counts
            3. CNV By Gene tab: Test row counts
            4. Fusion By Gene tab: Test row counts
            5. Fusion tab: Test row counts
        iii. OpenPedCan Gene Expression
            1. Linear boxplot: Test presence (T/F)

2. Log10 boxplot: Test presence (T/F)
3. Data download buttons: Test row counts

2. Disease Page
    a. Associated targets (indirect association): Test row count

3. Evidence Page
    a. OpenPedCan Somatic Alterations
        i. SNV By Gene tab: Test row counts
        ii. SNV By Variant tab: Test row counts
        iii. CNV By Gene tab: Test row counts
        iv. Fusion By Gene tab: Test row counts
        v. Fusion tab: Test row counts
    b. OpenPedCan Gene Expression
        i. Linear boxplot: Test presence (T/F)
        ii. Log10 boxplot: Test presence (T/F)
        iii. Data download buttons: Test row counts

4. FDA Pediatric Molecular Target Lists (FDA PMTL) page
    a. Test total row count
    b. Designation Relevant Molecular Target: Test row count
    c. Designation Non-Relevant Molecular Target: Test row count

5. Pediatric Cancer Data Navigation (PCDN) page
    a. Disease search results: Test row count
    b. Target search results: Test row count
    c. Target-Disease search results: Test datasource count