**Database Quality Control (QC) Report:** Testable PGx Genes.csv

**Date:** 10/02/2024  
**Prepared by:** Megha Jotangiya  
**Reviewed by:**

**1. Objective**

This Quality Control (QC) report aims to assess the accuracy, completeness, and reliability of data collected via web scraping for the [Database Name]. The methodology detailed below outlines the procedures used to verify the data integrity and ensure industry-standard quality for internal use.

**2. Methodology Overview**

**2.1 Scraping Process Validation**

* **Source Website:** <https://ddrx.pharmgkb.org/genotypes>
* **Scraping Method Used:** Python Script - [Build-Testable-PGx-Genes.ipynb](https://github.com/mjotangi/Testable-PGx-Genes/blob/main/Build%20Testable%20PGx%20Genes.ipynb)
* **Data Points Collected:** Gene, Genotype, Therapeutic Area(s) to be Tested, Associated Drug(s)
* **Frequency of Data Collection:** Daily

**2.2 Data Validation Techniques**  
The validation process consists of the following steps:

**2.2.1 Sample-Based Validation**

* **Sampling Rate:** 0.024739% of the 20211(records) scraped data was randomly selected against the original website for manual verification.
* **Sample Size:** 5 records/day is manually cross-checked.
* **Verification Method:** Manual comparison between scraped records and data on the source website.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PGx Gene** | **Genotype(s) to be Tested** | **Therapeutic Area(s)** | **Associated Drug(s)** | **Matches? With source website** | **Not Found Values** |
| CYP3A4 | \*17/\*22 | NERVOUS SYSTEM DRUGS | quetiapine | Yes | N/A |
| NUDT15 | \*3/\*4 | ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS | azathioprine, mercaptopurine, thioguanine | Yes | N/A |
| UGT1A1 | \*27/\*80+\*28 | ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS, ANTIINFECTIVES FOR SYSTEMIC USE | atazanavir, sacituzumab govitecan, nilotinib, dolutegravir, belinostat, pazopanib, raltegravir, irinotecan | Yes | N/A |
| CYP2B6 | \*18/\*5 | NERVOUS SYSTEM DRUGS, ANTIINFECTIVES FOR SYSTEMIC USE | sertraline, efavirenz | Yes | N/A |
| CFTR | 2789+5G->A/S549R(T>G) | RESPIRATORY SYSTEM DRUGS | ivacaftor | Yes | N/A |
| SLCO1B1 | \*5/\*8 | CARDIOVASCULAR SYSTEM DRUGS | pravastatin, simvastatin, rosuvastatin, pitavastatin, lovastatin, fluvastatin, atorvastatin | Yes | N/A |
| TPMT | \*10/\*3B | ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS | mercaptopurine, thioguanine, azathioprine | Yes | N/A |
| G6PD | Bangkok Noi | VARIOUS DRUG CLASSES IN ATC, NO GROUP ASSIGNED, MUSCULO-SKELETAL SYSTEM DRUGS, NERVOUS SYSTEM DRUGS, ALIMENTARY TRACT AND METABOLISM DRUGS, ANTIINFECTIVES FOR SYSTEMIC USE, ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS, ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS | toluidine blue, flutamide, Ascorbic acid (vitamin C), plain, mepivacaine, oxymetazoline and tetracaine, ceftriaxone, dabrafenib, sulfasalazine, nalidixic acid, sodium nitrite, articaine / epinephrine, pegloticase, moviprep, bupivacaine, dapsone, primaquine, tafenoquine, glyburide, rasburicase, lidocaine / prilocaine, glipizide, lidocaine and tetracaine, glimepiride, chloroquine, chlorpropamide, hydroxychloroquine, nitrofurantoin, tolazamide, methylene blue, tolbutamide | Yes | N/A |
| DPYD | c.1108A>G/c.295\_298delTCAT (\*7) | ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS | fluorouracil, capecitabine | Yes | N/A |
| CYP2C9 | \*10/\*2 | BLOOD AND BLOOD FORMING ORGAN DRUGS | warfarin | Yes | N/A |
| CYP2D6 | \*2x‚â•3/\*32 | RESPIRATORY SYSTEM DRUGS, ALIMENTARY TRACT AND METABOLISM DRUGS, ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS, NERVOUS SYSTEM DRUGS | tropisetron, paroxetine, atomoxetine, desipramine | No | vortioxetine, trimipramine, tramadol, tamoxifen, ondansetron, nortriptyline, meclizine, imipramine, eliglustat, doxepin, codeine, clomipramine, amoxapine, amitriptyline, acetaminophen / caffeine / dihydrocodeine |
| RYR1 | c.467G>A/c.7063C>T | MUSCULO-SKELETAL SYSTEM DRUGS, NERVOUS SYSTEM DRUGS | succinylcholine, halothane | No | sevoflurane, methoxyflurane, isoflurane, enflurane, desflurane, |
| G6PD | B (reference)/Crispim | VARIOUS DRUG CLASSES IN ATC, NO GROUP ASSIGNED, MUSCULO-SKELETAL SYSTEM DRUGS, NERVOUS SYSTEM DRUGS, ALIMENTARY TRACT AND METABOLISM DRUGS, ANTIINFECTIVES FOR SYSTEMIC USE, ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS, ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS | toluidine blue, flutamide, Ascorbic acid (vitamin C), plain, mepivacaine, oxymetazoline and tetracaine, ceftriaxone, dabrafenib, sulfasalazine, nalidixic acid, sodium nitrite, articaine / epinephrine, pegloticase, moviprep, bupivacaine, dapsone, primaquine, tafenoquine, glyburide, rasburicase, lidocaine / prilocaine, glipizide, lidocaine and tetracaine, glimepiride, chloroquine, chlorpropamide, hydroxychloroquine, nitrofurantoin, tolazamide, methylene blue, tolbutamide | Yes | N/A |

**2.2.2 Automated Integrity Checks**

* **Data Type Validation:** Ensures the data types in the database (e.g., text, numerical values, dates) match the expected formats.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Field** | **Expected Format** | **Actual Format** | **Status** | **Notes** |
| **PGx Gene** | Text | TRUE | Corrected | Used =ISTEXT() |
| **Genotype(s)** | Text | TRUE | Corrected | Used =ISTEXT() |
| **Therapeutic Area(s)** | Text | TRUE | Corrected | Used =ISTEXT() |
| **Associated Drug(s)** | Text | TRUE | Corrected | Used =ISTEXT() |

* **Inconsistency Correction:**
  + Proper cases are applied to therapeutic areas and drugs (e.g., "alimentary tract and metabolism drugs" to "Alimentary Tract and Metabolism Drugs").
  + **Correction Method:** Applied =PROPER() formula to affected rows.
* **Uniqueness Checks:** Verifies that key identifiers (e.g., product IDs, URLs) remain unique and free of duplicates.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **Duplicate Genotype** | **Status** | **Remark** |
| CYP2C9 | \*15/\*27 | Removed | Manually Removed using pivot table |
| CYP2C9 | \*15/\*59 | Removed | Manually Removed using pivot table |
| CYP2C9 | \*40/\*54 | Removed | Manually Removed using pivot table |
| G6PD | Alhambra/Telti, Kobe | Removed | Manually Removed using pivot table |
| G6PD | Ananindeua/Lille | Removed | Manually Removed using pivot table |
| G6PD | Aures/Ierapetra | Removed | Manually Removed using pivot table |
| TPMT | \*4/\*5 | Removed | Manually Removed using pivot table |

* **Null Value Identification:** Flags records where key data fields are missing or contain null values.
* **Range & Consistency Check:** No numerical data.

**2.2.3 Consistency with Historical Data**

* **Change Detection:** Cross-references newly scraped data against historical records to identify unexpected changes or outliers.
* **Versioning Review:** Scraped data will be compared with prior versions to identify potential misalignment or degradation in data quality.

**3. Data Quality Metrics**

We define the following data quality metrics as per industry standards:

| **Metric** | **Target (%)** | **Achieved (%)** | **Remarks** |
| --- | --- | --- | --- |
| **Data Accuracy** | 99.5% | 95% | Discrepancies in Associated Drug(s) and Therapeutic Area(s) column for the gene that has more records. |
| **Data Completeness** | 99.0% | 95% | All fields are expected to be fill correctly. |
| **Data Consistency** | 98.5% | 98.8% | Formatting issues resolved. |
| **Uniqueness** | 100% | 100% | Duplicates removed. |
| **Timeliness** | 95% | 95% | Scraping occurred for each gene one by one |

**4. Results**

**4.1 Sample Verification Results**

* **Records Verified:** 5 per day
* **Discrepancies Found:** Discrepancies found in Associated Drug(s) and Therapeutic Area(s) column
* **Discrepancy Breakdown:** Missing drugs for certain genotypes, as found in manual verification.

**4.2 Automated Integrity Check Results**

* **Errors Detected:** 6 duplicate entries were identified and removed.
* **Resolution:** Duplicates were manually removed, and text formatting errors were corrected.

**4.3 Consistency Check Results**

* **Inconsistencies Identified:** 2 discrepancies regarding missing drugs for *CYP2D6* and *RYR1*.
* **Resolution Approach:** Rescraping

**5. Conclusion & Recommendations**

**5.1 Conclusion**  
Based on the sample verification, the database scraping is functioning well. The data is largely accurate, complete, and consistent, with minor corrected discrepancies.

**5.2 Recommendations**

* **For Ongoing Quality Control:**
* Continue sample-based validation to monitor accuracy.
* Increase the sample size for manual verification to ensure accuracy further.
* Implement real-time alerts for discrepancies in data scraping results.
* **Improvements in the Scraping Process:**
* Introduce automated drug list validation to minimize missing drug and therapeutic area data.
* Improve scraping logic to avoid missing rare drug entries.
* Monitor PharmGKB's structure for potential changes that might affect future scraping results.

**6. Appendix**

* **Raw Data Samples:** N/A
* **Error Logs:** N/A
* **Automated Script Reports:** The data is manually checked.