Developing a FHIR-based Framework for Phenome Wide Association Studies: A Case Study with A Pan-Cancer Cohort

Nansu Zong, Ph.D. 1, Deepak K. Sharma, M.S.1, Yue Yu, Ph.D. 1, Jan B. Egan, Ph.D. 2, Jaime I. Davila, Ph.D. 1, Chen Wang, Ph.D. 19, Guoqian Jiang, M.D., Ph.D. 19
1 Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
2 Center for Individualized Medicine, Mayo Clinic, Rochester, MN

Abstract

Phenome Wide Association Studies (PheWAS) enables phenome-wide scans to discover novel associations between genotype and clinical phenotypes via linking available genomic reports and large-scale Electronic Health Record (EHR). Data heterogeneity from different EHR systems and genetic reports has been a critical challenge that hinders meaningful validation. To address this, we propose an FHIR-based framework to model the PheWAS study in a standard manner. We developed an FHIR-based data model profile to enable the standard representation of data elements from genetic reports and EHR data that are used in the PheWAS study. As a proof-of-concept, we implemented the proposed method using a cohort of 1,595 pan-cancer patients with genetic reports from Foundation Medicine as well as the corresponding lab tests and diagnosis from Mayo EHRs. A PheWAS study is conducted and 81 significant genotype-phenotype associations are identified, in which 36 significant associations for cancers are validated based on a literature review.

1. Introduction

Genome-wide association studies (GWAS) performs a broad range of statistical tests over participants with disease and without the disease to investigate the relation between the genome-wide set of genetic alterations and the disease (1). While thousands of SNPs have been successfully identified to have associations with diseases (e.g., type 2 diabetes and osteoporosis (2)), GWAS only focuses on a restricted phenotypic domain based on the existence or nonexistence of a particular disease, and neglects the benefit of using of phenotypes (including sub-phenotypes, biomarkers, and endophenotypes). As genomic information and investment into the development of large-scale Electronic Health Record (EHR) systems grow, a new trend to link rich phenotypic data accessible in EHR systems to genomic data indicates a supplementary alternative, Phenome-wide association study (PheWAS) (3, 4). PheWAS offers phenome-wide scans for exploring new associations between genotype and clinical phenotypes and gives insights into underlying biological associations. PheWAS has been effectively employed as a proof-of-concept to discover both expected associations (3, 5) as well as possible new associations (6).

The main idea behind PheWAS is to connect genetic data with phenotypic data in a variety of medical systems to discover the genotype-phenotype associations. Integrating heterogeneous medical records with genetic data can provide quantitative measurements (e.g., laboratory tests) along with detailed disease conditions (e.g., diagnosis), which increases power for statistical analysis. However, data heterogeneity is commonly encountered when integrating data from multiple sources thus hindering EHR data usage (7). In addition, PheWAS studies are limited by data biases in single-center research (8, 9). Those issues can be solved by the validation of PheWAS conducted across multiple institutions, which utilize heterogeneous medical records. Population Architecture using Genomics and Epidemiology (PAGE) network is an example of such endeavors (5, 8).

A benefit of having a standard input data format in addition to having a clear standardized definition of what data is present and what values are acceptable is plug and play functionality for conducting any PheWAS studies enabling replication across the different institutions so long as data is generated in that standardized format. The model-driven approach for standardizing phenotypic data has been increasingly adopted by the cancer research informatics community. For example, cancer profiles for clinical applications, such as breast cancer, colorectal cancer, prostate cancer, have been developed by the Clinical Data Interchange Standards Consortium (CDISC) (10), Clinical Information Modeling Initiative (CIMI) (11), and Royal College of Pathologists of Australasia (RCPA) (12). Nevertheless, such a profile that represents patient genotypes (e.g., from the genetic report) and phenotypes (e.g., from lab test and diagnosis) for PheWAS study have yet to be studied. Furthermore, to support precision medicine, the promotion of the standard to facilitate exchanging of data, such as clinical and genomic information, between parties is the primary focus in healthcare communities, such as HL7 Clinical Genomics (13). There are many standardized data models, including Informatics for Integrating Biology and the Bedside (i2b2) (14), the National
Quality Forum (NQF) Quality Data Model (QDM) (15), the OHDSI Common Data Model (CDM) (16), the HL7 Consolidated Clinical Document Architecture (CDA) (17). The Fast Healthcare Interoperability Resources (FHIR) enables the quickly exchanging of EHR data, which is considered as a next-generation standards framework (18). FHIR is widely adopted by the major modern EHR vendors (e.g., Epic) and healthcare providers (e.g., Mayo Clinic and Intermountain Healthcare) (19).

By creating a standardized data model profile to facilitate the PheWAS study across diverse research institutions, we proposed a framework to populate the data of a PheWAS in a standard manner. We developed an FHIR-based data model profile to enable the standard representation of the data elements used in the PheWAS studies. As a proof-of-concept, we implemented the proposed method and the data model profile using 1,595 genetic reports from Foundation Medicine as well as the corresponding lab tests and diagnosis codes from Mayo Clinic’s clinical data warehouse known as the Unified Data Platform (UDP). An aggregated PheWAS study was conducted based on the data matrices generated from the proposed data profile, and 81 significant genotype-phenotype associations were identified as results, in which 36 significant associations for cancers were validated based on a literature review.

2. Methods

The framework is designed to enable the results of PheWAS to be validated across multiple organizations via the adoption of an FHIR-based data profile. There are mainly three modules in the framework as shown in Figure 1: 1) data preparation and preprocessing, where the data is generated from diverse EHR systems and databases, 2) FHIR-based data model profiling, where an FHIR-based profile is developed based on the PheWAS study criteria, 3) mapping to FHIR-based data profile, where the local data schema will be mapped to the standardized profile, and 4) data population and PheWAS, where the local data will be used to populate the matrices for PheWAS based on the FHIR-based profile.

Figure 1. The framework of FHIR-based PheWAS study

2.1. Data preparation and preprocessing

The two sources, genetic reports, and EHR data are used in this study. For the genetic reports, we utilized the 1,595 reports generated from Foundation Medicine, which is a clinically available test that provides actionable information based on the results of the individual genomic profile of each patient’s cancer. Every test result provides microsatellite instability (MSI) and tumor mutational burden (TMB) to assist immunotherapy decisions. For the diagnosis and lab tests, we extracted the EHR data from Mayo Clinic’s UDP (20). The UDP is a clinical data warehouse that provides a combined view of multiple heterogeneous data across multiple databases, e.g., EPIC-based EHR. To integrate genetic reports and EHR data, we mapped the patients based on three data elements: 1) patient clinic number, 2) names (first and last name), and 3) Date Of Birth (DOB). In practice, if only the names and DOB were matched for a patient, a manual review was conducted for accurate mapping. Even though such a method provides the mappings with high precision and recall rates, it may not be feasible on larger datasets. A customized matching strategy that uses more features, such as race, sex, zip code, may provide an automated solution with an acceptable recall rate in other cases. The report issue time was also recorded for the population of the matrix in

751
Section 2.4. The diagnosis and lab tests are extracted from EHR based on the mapped patients. For the diagnosis, all diagnosed diseases across the entirety of visits were collected. The diseases were encoded with International Classification of Disease (ICD-9/10) codes and Phecode (3), which is a custom grouping of ICD9 billing codes to approximate the clinical disease phenotype. Similarly, lab test records were collected from all visits. The Logical Observation Identifiers Names and Codes (LOINC) was adopted to encode the lab test items. The values were normalized to remove the noises, e.g., “Neg”, “N”, and “Negative” are represented with “negative”. The top 10 elements in each dataset can be found in Table 1.

Table 1. Distribution of the top 10 elements in each report.

<table>
<thead>
<tr>
<th>ID</th>
<th>Genes</th>
<th>Diagnosis</th>
<th>Lab tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
<td># Records (Percentage)</td>
<td>PheneCode “Description”</td>
</tr>
<tr>
<td>1</td>
<td>TP53</td>
<td>818 (51.29%)</td>
<td>198 &quot;Secondary malignant neoplasm&quot;</td>
</tr>
<tr>
<td>2</td>
<td>KRAS</td>
<td>389 (24.39%)</td>
<td>401 &quot;Hypertension&quot;</td>
</tr>
<tr>
<td>3</td>
<td>CDKN2A/B</td>
<td>194 (12.16%)</td>
<td>401.1 &quot;Essential hypertension&quot;</td>
</tr>
<tr>
<td>4</td>
<td>PIK3CA</td>
<td>165 (10.34%)</td>
<td>272 &quot;Disorders of lipid metabolism&quot;</td>
</tr>
<tr>
<td>5</td>
<td>APC</td>
<td>155 (9.72%)</td>
<td>272.1 &quot;Hyperlipidemia&quot;</td>
</tr>
<tr>
<td>6</td>
<td>PTEN</td>
<td>143 (8.89%)</td>
<td>285 &quot;Other anemias&quot;</td>
</tr>
<tr>
<td>7</td>
<td>ARID1A</td>
<td>131 (8.21%)</td>
<td>427 &quot;Cardiac dysrhythmias&quot;</td>
</tr>
<tr>
<td>8</td>
<td>CDKN2A</td>
<td>126 (7.90%)</td>
<td>512 &quot;Other symptoms of respiratory system&quot;</td>
</tr>
<tr>
<td>9</td>
<td>RB1</td>
<td>120 (7.52%)</td>
<td>276 &quot;Disorders of fluid&quot;</td>
</tr>
<tr>
<td>10</td>
<td>TERT</td>
<td>111 (6.96%)</td>
<td>198.1 &quot;Secondary malignancy of lymph nodes&quot;</td>
</tr>
</tbody>
</table>

2.2. FHIR-based data model profile

We proposed a data model profile using the FHIR modeling mechanism to functionalize the standard representation of genetic reports, lab test results, and diagnosis for the PheWAS study. The genetic entries were modeled by the developed profile, “PheWASGeneticReport”. This profile was derived from the existing profile “Observation-genetics”, which was generated based on the resources “Observation”. “PheWASGeneticReport” models observations about a mutated gene based on the extension “Observation-geneticsGene”. The alterations were modeled with the extension, “Observation-geneticsVariant”, where the three types of alternations were encoded with, 1) Disease-relevant genomic alteration, 2) Variants of Unknown Significance (VUS), and 3) Disease-relevant gene with no reportable alterations identified. The lab test entries were modeled by the developed profile, “PheWASLabTest”, based on the resource “Observation”. “PheWASLabTest” specifies the constraint with ‘code’ and ‘value’ to model the test items. A test item was encoded with LOINC with three types of the scale used in PheWAS: numeric, ordinal, and categorical. Of note, for the Quantitative (Qn) type in LOINC, relational operators (e.g., <, >, and =) were removed to change into numeric or categorized (e.g., 1-10). We have defined standard units of measurement to be used for normalizing numeric values. However, selecting custom units for localized studies is also supported. If the results are the only element that requires validation across different organizations, from a statistical perspective, using consistent units of measurement within the organization is sufficient for conducting a localized PheWAS study. The types, Narratives (Nar), Multi, Document (Doc), and Set were ignored. We modeled the diagnosis entries with multiple diagnosis reports across the entirety of visits. While the lab test entities only modeled the latest visit regarding the issue date of the genetic report, the profile could be easily extended by adding more elements to represent multiple values, e.g., minimum, maximum, mean, and median, for the entries from multiple visits. The diagnosis and lab test values for each visit can also be recorded based on the same extension mechanism.

The diagnosis entries were modeled by “Condition” with a developed extension, “PheWASDiagnosis”, where each disease was encoded with ICD 9/10 and Phecode vocabularies along with a frequency counter. The logical model in UML can be found in Figure 2.
Figure 2. FHIR-based PheWAS data model profile based on FHIR resources.

In practice, we used FHIR Release 4 (R4) (21) for laying out the model elements. In the absence of a UML symbol for the profile, we reused the generalization symbol (inheritance) and distinguished profiles from extensions by showing the class namespaces. The PheWAS model extensions and profiles were created using the Forge editor (22). The UML Model was put together by manually extending and profiling the imported FHIR entities. A detailed model report document and its web rendering are available at (https://github.com/BD2KOnFH/phewas-on-fhir).

2.3. Mapping to FHIR-based data model profile

Figure 3. The mapping between PheWAS patient profile and the FHIR-based PheWAS data model profile. The elements of PheWAS patient profile are in green. The FHIR resources are represented in purple and the items are in blue. The newly generated items based on FHIR “Extension” are in yellow.

To populate the data for PheWAS with FHIR resources, we established a mapping between the FHIR resources and local data as shown in Figure 3. For the general information, the “identifiers”, “status”, and “subjects”, “cohort”, and
“reported” are mapped to the corresponding elements in “Condition”. For each genetic report of a patient, the profile “PheWASGeneticReport” was mapped. Specifically, the item “tumorType” was mapped to “bodySite” and the remaining items mapped to the corresponding defined extensions. Of note, different with the cardinality of “gene” (1..*) and “variant” (1..*) in the genetic report, “PheWASGeneticReport” has a cardinality defined as “0..1” in “Observation”, therefore, a data entry of the genetic report will be represented with multiple data entries based on “PheWASGeneticReport”. The lab test entries were mapped to “PheWASLabTest”, where “code”, “value”, and “unit” were mapped correspondingly. The diagnosis entities were mapped to “PheWASDiagnosis”, where “Count” and “ICD9”/“Phecode” were mapped correspondingly.

2.4. Data population and PheWAS

Three matrices were populated based on the patient profile modeled in Section 2.2, which are patient-genetic, patient-lab test, and patient-diagnosis. To form the matrices, the elements in diagnosis and lab test (i.e., diseases for patient-diagnosis and test item for patient-lab test matrix), and the elements (i.e., genes for patient-genetic matrix) in the genetic report, were extracted respectively as the columns. Each patient record was considered as a row in the matrices. For a patient in the patient-genetic and patient-diagnosis matrix, each cell indicated the presence/absence of each reported gene variant and disease diagnosis. For a patient in the genetic-lab test matrix, each cell is the value of each lab test.

We conducted two kinds of tests based on two sets of cohorts that corresponded to the three matrices in an aggregated PheWAS for gene mutations. For patient-diagnosis cohorts, a case was a patient record with a valid phocode while other records were labeled as a control. We calculated the case and control chi-square distribution- associated allelic p-value. We selected only those that occurred in a minimum of 10 cases as a threshold of clinical interest. For genetic-lab test cohorts, since all the lab test variables in this study are numerical, we conducted the Kolmogorov–Smirnov (KS) test of the value distributions for each gene-lab test pair. If the lab test with the observed cell counts fell below a 10% threshold, then the entire population was filtered out of the study. Since the conventional Bonferroni correction is considered conservative for PheWAS (3, 4, 23), we adjusted all the p-values by FDR (24).

In addition, we also conducted a literature review to validate whether a significant association identified from the PheWAS studies was a known association or not.

3. Experiment and Results

Based on the two kinds of data matrices, we conducted two PheWAS studies, gene v.s. lab test and gene v.s. phocode, which revealed well-established associations with significant p-values. As shown in Figure 4, four associations were identified between the genes - CDKN2A/B, CDKN2A, TERT, and SKT11, and the lab tests. CDKN2A/B was found significantly related to blood monocytes count (p-value=0.0269 in Figure 4) as it is related to the regulation of monocyte–macrophage function. For example, CDKN2A/B locus performs as a modifier on atherosclerosis. Atherogenesis is enhanced by the transplantation of heterozygous CDKN2A-deficient bone marrow,

Figure 4. Heatmap of the correlation of genes and lab tests.
which increases the circulation of pro-inflammatory Ly6Chi monocytes and proliferation of peritoneal monocyte/macrophage (25, 26). CNKD2A is also responsible for the activation of the D-CDK4/6-INK4-Rb pathway. The pharmacodynamic decreasing in neutrophil counts is related to the increase of palbociclib exposure, which is a treatment for CNKD2A mutation (27). Our study illustrated the correlation between CNKD2A and neutrophil counts in blood (p-value=0.003). In addition, neutrophil counts are also correlated to STK11 (p-value=0.0269), which is the most commonly inactivated tumor. Genetic ablation of STK11/LKB1 results in the accumulation of neutrophils in non-small cell lung cancer (NSCLC) (28, 29). Myeloproliferative neoplasms (MPN) are a group of diseases, which produce excess cells in the bone marrow. They can develop myelodysplastic syndrome and acute myeloid leukemia. We showed the correlation between TERT and erythrocyte count in blood (p-value=0.0379), where TERT mutations increase the proliferation of common myeloid progenitor to affect hematopoiesis (30, 31).

**Figure 5.** Heatmap of the correlation of genes and diagnosis (phecode).

We identified 58 significant associations between the genes and phecodes shown in Figure 5. Top 5 genes that have the most phecode-related associations are KRAS (9 associations), STK11 (7 associations), TP53 (5 associations), APC (5 associations), and BRAF (5 associations). The detail of the frequency for # association is shown in Figure 5. We further validated 32 associations for the genes related to the carcinomas by a literature review, in which KRAS, APC, TP53 are the genes having the most cancer-related associations. The rest of the associations, which are potentially novel, are listed in Table 3.

**Table 3.** Potential correlations without validation.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Diagnosis</th>
<th>P-value</th>
<th>Gene</th>
<th>Diagnosis</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>APC</td>
<td>Ileostomy status</td>
<td>5.49E-05</td>
<td>MLL2</td>
<td>Bone marrow or stem cell transplant</td>
<td>1.70E-02</td>
</tr>
<tr>
<td>BRAF</td>
<td>Diseases of the larynx and vocal cords</td>
<td>9.87E-06</td>
<td>MLL2</td>
<td>Herpes simplex</td>
<td>1.70E-02</td>
</tr>
<tr>
<td>BRAF</td>
<td>Nontoxic multinodular goiter</td>
<td>2.30E-05</td>
<td>MLL2</td>
<td>Non-Hodgkins lymphoma</td>
<td>1.21E-12</td>
</tr>
<tr>
<td>BRAF</td>
<td>Nontoxic nodular goiter</td>
<td>2.04E-04</td>
<td>STK11</td>
<td>Degenerative skin conditions and other dermatoses</td>
<td>1.60E-02</td>
</tr>
<tr>
<td>BRAF</td>
<td>Secondary hypothyroidism</td>
<td>9.55E-05</td>
<td>STK11</td>
<td>Dizziness and giddiness (Light-headedness and vertigo)</td>
<td>3.48E-02</td>
</tr>
<tr>
<td>CCND1</td>
<td>Acquired absence of breast</td>
<td>3.69E-03</td>
<td>TERT</td>
<td>Chronic airway obstruction</td>
<td>3.74E-03</td>
</tr>
<tr>
<td>CCNE1</td>
<td>Cancer of other female genital organs</td>
<td>1.02E-06</td>
<td>STK11</td>
<td>Emphysema</td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td>Postmenopausal atrophic vaginitis</td>
<td>4.86E-02</td>
<td>STK11</td>
<td>Keratoderma, acquired</td>
<td>8.44E-03</td>
</tr>
<tr>
<td>KRAS</td>
<td>Diseases of pancreas</td>
<td>3.95E-04</td>
<td>STK11</td>
<td>Peripheral or central vertigo</td>
<td>5.47E-05</td>
</tr>
<tr>
<td>KRAS</td>
<td>Obstruction of bile duct</td>
<td>1.05E-03</td>
<td>STK11</td>
<td>Vertiginous syndromes and other disorders of vestibular system</td>
<td>1.78E-03</td>
</tr>
<tr>
<td>KRAS</td>
<td>Other biliary tract disease</td>
<td>5.87E-04</td>
<td>TERT</td>
<td>Nontoxic multinodular goiter</td>
<td>9.75E-05</td>
</tr>
<tr>
<td>KRAS</td>
<td>Other disorders of biliary tract</td>
<td>2.59E-02</td>
<td>TERT</td>
<td>Secondary hypothyroidism</td>
<td>7.77E-06</td>
</tr>
<tr>
<td>MYC</td>
<td>Encounter for long-term (current) use of antibiotics</td>
<td>1.04E-02</td>
<td>TERT</td>
<td>Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]</td>
<td>1.45E-02</td>
</tr>
</tbody>
</table>
KRAS functions as a switch for cell signaling and controlling cell proliferation during its normal function. However, when KRAS mutates, it will disrupt negative signaling and cause cells to proliferate and grow into cancer. The effect of KRAS mutations depends on the order of the mutations. If the KRAS mutation occur after APC mutation, it often develops into cancers, such as colorectal (p-value=6.62 E-07), colon (p-value=1.65 E-04), and rectum cancers (p-value=4.51 E-04) (32, 33).

In addition, somatic KRAS mutations are commonly found in pancreatic cancer (p-value=4.51E-04) (9.57 E-16) (34). According to a previous study (35), the most frequent metastatic sites of 468 lung adenocarcinoma patients were lung (45.6%), bone (26.2%) (p-value= 1.12 E-02), adrenal gland (17.4%), brain (16.8%), pleura (15.6%) and liver (11%). APC is regarded as a tumor suppressor gene as it could prevent uncontrolled cell growth that can lead to cancerous tumors. The protein encoded by the APC gene plays a key role in determining whether or not a cell will grow into a tumor. APC mutation and inactivation is a critical event to malignant rectum neoplasm (p-value=6.9 E-23), colorectal (p-value=2.04 E-44) and colon tumorigenesis (p-value=1.24 E-29) (36-38). In addition, the liver metastasis (p-value= 1.57 E-04) could also be caused by the combinations including APC, KRAS, and TGFB2 mutations (39). TP53 encodes the tumor protein p53, which is critical for tumor suppressor in multicellular organisms. TP53 is the most prevalent mutated gene in human cancers (> 50%), implying that the TP53 gene plays a vital role in the prevention of cancer formation. The accumulation of genetic mutations in the driving genes contributes to colorectal cancer development and malignant progression. APC, KRAS, and TP53 (p-value= 2.6 E-02) are often observed as driver genes. TP53 mutations, for instance, are discovered in 60% of colorectal cancers (40, 41). TP53 is also regarded as the significant genetic variant of human ovarian epithelial and genital cancer (p-value=1.45E-02 for malignant neoplasm of the ovary, p-value=1.03 E-02 for malignant neoplasm of the ovary and other uterine adnexa, and p-value= 3.47E-03 for cancer of other female genital organs) (42-44). Multiple cancer metastases are caused by TP53, such as metastasis of gastrointestinal cancer (p-value=1.45E-02 for secondary malignant neoplasm of digestive systems) (45).

4. Discussion and conclusion

To facilitate the validation of PheWAS-based studies across different research organizations, we proposed an FHIR-based PheWAS data model profile to enable the standard representation of the data elements from genetic reports and EHR data that are used in the PheWAS study. A framework to automate the data population for PheWAS was introduced. As a proof-of-concept, we implemented the proposed method based on 1,595 genetic reports from FoundationOne CDx as well as the corresponding lab tests and diagnosis from Mayo Clinic’s UDP. A PheWAS study was conducted and 81 significant genotype-phenotype associations were obtained as a result. We have validated 36 significant genetic mutations for cancers based on a literature review.

There are several significant contributions and advantages of this study. Firstly, we demonstrated that it is feasible to represent the PheWAS study data using FHIR. To the best of our knowledge, this is the first study to apply standardization to model PheWAS study, with the ultimate goal of facilitating cross-validation for PheWAS studies. Secondly, we developed an FHIR-based data model profile that represents the data elements needed for PheWAS. The model uses the resources and profiles from FHIR, where a number of open-source validation mechanisms and tools, such as FHIR specification and implementation guides, are supported by the FHIR community for ensuring data quality. Our data model profile adopts the FHIR specifications to enable the modification of the constraints and rules to accommodate real data, which can be easily adapted and extended.

There are a number of limitations in this study we would like to tackle for the future work. First, the genetic reports are from FoundationOne CDx. As FoundationOne did not provide information on whether the data is generated from the tumor or normal samples, we were unable to separate germline mutations from somatic mutations, which values differ in initial diagnosis and progression, and are critical for cancer studies. The failure to capture differences in genetic data weakens our contribution to cancer studies, which is considered as a limitation of this study. Since the datasets used in this study are from multiple sources, an extra mapping effort between the proposed data profile and the schema of the local datasets is needed to enable the data population. Nevertheless, with the FHIR-based APIs under development through HL7 Argonaut project (46), such mapping between the local data and the proposed data profile will no longer be needed, which greatly promotes the flexibility and adaptability of the proposed framework. Second, by further exploring dependent phenotypes related to the same genetic alteration (e.g. KRAS and colorectal cancer v.s. KRAS and malignant neoplasm of rectum, rectosigmoid junction, and anus), we notice the limitations of performing individual genetics-phenotype associations without taking into consideration phenotype dependence and ontology structure. The biological classification of the phenotypes illustrates a hidden connection between cancers. Therefore, a more sophisticated PheWAS methodology can be designed to leverage genetic and phenotype
ontologies structure to enhance the power of discoveries. In addition, by mapping lab tests to LOINC codes to more comprehensive Human Phenotype Ontology (HPO) (47, 48), will greatly advance the design of the refinement methodology for PheWAS. Third, the ultimate objective of the proposed method is to facilitate validation for PheWAS studies across multiple organizations. Due to a lack of resources, we cannot demonstrate the use case in this study. We plan to reach out to other institutions or research networks (eg, eMERGE research network (49)) for conducting PheWAS studies based on the proposed framework to have a comprehensive evaluation. In addition, for the potential new associations identified (see Table 3), further validation can also be conducted. Fourth, this study only considers the values of the diagnosis and lab test for the most recent visit. However, since the diagnosis and lab test values may change over time, the temporal aspect needs to be taken into consideration for building more sophisticated models for enabling unconfounded findings. Our future work will develop such a model based on the proposed framework. Fifth, our PheWAS study is carried out on common variants. It will be valuable to explore how the proposed FHIR-based framework will perform with common variants and non-cancer phenotypes in our future work. Lastly, the proposed method developed based on the data standard, FHIR, was selected for two major reasons: 1) FHIR is widely adopted among all modern EHR vendors and data providers and can be easily adopted. The adoption requires less Extract Transform Load (ETL) effort for the data representation from original data sources with the proposed data model; and 2) FHIR is not a data standardization model for data storage and management but rather a data communication method for efficiently exchanging medical data among organizations, which fits the original purpose of facilitating the validation for PheWAS studies across different organizations. Although there are benefits, as mentioned above, regarding the adoption of FHIR for PheWAS, we agree that the adoption of other standardization data models, such as OHDSI CDM, can be necessary for some cases where data storage and management system is needed for long-term scientific needs. For such cases, comprehensive data modeling strategies based on diverse standardization models require further study in the future.

Acknowledgments

This study is supported by the NIH BD2K grant U01 HG009450 and the Center for Individualized Medicine at Mayo Clinic.

References

22. FORGE.