Evaluation of the PharmGKB Knowledge Base as a Resource for Efficiently Assessing the Clinical Validity and Utility of Pharmacogenetic Assays

Kensaku Kawamoto, MD, PhD1,3, Lori A. Orlando, MD, MHS2, Deepak Voora, MD2, David F. Lobach, MD, PhD, MS1, Scott Joy, MD2,3, Alex Cho, MD, MBA2,3, Geoffrey S. Ginsburg, MD, PhD2,3
1Division of Clinical Informatics, Department of Community and Family Medicine; 2Department of Medicine; and 3Duke Institute for Genome Sciences & Policy, Duke University, Durham, NC

ABSTRACT

Prior to clinical use, pharmacogenetic tests should be systematically evaluated for their clinical validity and utility. Here, we evaluated whether the publicly available, online Pharmacogenomics Knowledge Base (PharmGKB) could facilitate such assessments by efficiently identifying relevant peer-reviewed manuscripts. The search targets were 55 manuscripts regarding clinical validity and utility included in systematic reviews of warfarin, antidepressant, and irinotecan pharmacogenetics. When direct inclusion in PharmGKB was the search criterion, recall was 33% and precision was 16%. However, recall increased to 78% when citation within a PharmGKB-identified manuscript was added as a search criterion. These recalled manuscripts accounted for 87% of the study subjects, and domain experts determined that the omission of the remaining manuscripts was unlikely to have changed the conclusions of the reviews. Thus, we conclude that PharmGKB can facilitate the systematic assessment of pharmacogenetic assays through the efficient identification of relevant peer-reviewed manuscripts.

INTRODUCTION

Genetic variation can significantly influence how individuals respond to drug therapies.1 The study of this phenomenon, referred to as pharmacogenetics when one gene is involved and pharmacogenomics when multiple genes are involved, has revealed numerous relationships between individuals’ genotypes and their responses to drug therapies. Currently, over 500 drugs have been evaluated for their differential effects based on genotype.2

The use of pharmacogenetic (PGx) testing could significantly improve health and healthcare. For example, genetically guided drug therapy could limit the need for using a trial-and-error approach to identify drug therapies that are effective for a given patient; increase medication adherence, which is estimated to be only 50% for chronic diseases5; and decrease adverse drug reactions, which lead to over 100,000 deaths annually in the United States.4

Prior to clinical use, however, PGx tests should be systematically evaluated with regard to clinical validity (the ability of the test to detect or predict the disorder/phenotype of interest) and clinical utility (the risks and benefits associated with the introduction of the test into practice).5 Also, a genetic test should be analytically valid (i.e., be able to measure the genotype accurately and reliably).5

Two rigorous approaches to evaluating the analytic validity, clinical validity, and clinical utility of genetic assays are the EGAPP (Evaluation of Genomic Applications in Practice and Prevention)6 and ACCE (Analytic validity, Clinical validity, Clinical utility, and Ethical, legal, and social implications)5 review methodologies. The ACCE initiative was launched by the Centers for Disease Control and Prevention (CDC) in 2000, and EGAPP was launched by the CDC in 2004 to build on the ACCE methodology. A streamlined version of ACCE known as rapid ACCE is also available.7

While rigorous, these approaches to evaluating genetic assays have been limited by their significant resource requirements. Full ACCE reviews can cost $50,000 to $100,000 per test, and even rapid ACCE reviews can cost up to $40,000.7 Consequently, to our knowledge, less than a dozen reviews have been published using these methodologies to date.

An important contributor to the high cost of these assessments is the effort required to systematically identify relevant manuscripts. For example, a 2007 EGAPP evaluation of cytochrome P450 genotyping to guide the use of selective serotonin reuptake inhibitors (SSRIs) required the screening of 1,200 abstracts to identify less than 40 relevant manuscripts.8 Similarly, a 2009 EGAPP evaluation of the use of UGT1A1 genotyping to guide the treatment of metastatic colorectal cancer with irinotecan involved the screening of over 2,400 abstracts to identify less than 30 relevant articles.9

A potential resource for facilitating the identification of relevant manuscripts for such systematic assessments is the Pharmacogenomics Knowledge Base (PharmGKB).2 Available on the Web at www.pharmgkb.org, PharmGKB is a publicly available tool developed by Stanford University through funding from the National Institutes of Health. Through the use of automated literature analyses, submissions by affiliated research scientists, and manual curation by a team of seven primarily Ph.D.-level scientists, PharmGKB provides expertly curated information on the relationships between over 500 drugs, 1200 genetic variants, and 500 diseases.10 Of note, associations between drugs
and genetic variants are justified by references to manuscripts which can be retrieved through the Web site as well as through a Web service interface.

While PharmGKB is primarily intended to facilitate basic research, we hypothesized that PharmGKB could also facilitate the systematic assessment of the clinical value of PGx tests. In particular, we hypothesized that PharmGKB could be used to efficiently identify peer-reviewed manuscripts important for assessing the clinical validity and utility of PGx tests. Thus, we selected three ACCE and EGAPP reviews of PGx tests and evaluated whether PharmGKB could be used to identify the manuscripts included as the primary data sources in these gold standard reviews.

METHODS

Overview. The study methods are summarized in Figure 1. First, three gold standard systematic reviews of PGx testing were identified. Second, the manuscripts used as the primary data sources in these reviews were abstracted and classified. Third, PharmGKB was searched for relevant manuscript entries. Fourth, we conducted citation analyses to evaluate whether targeted manuscripts were cited by PharmGKB manuscripts. Fifth, recall and precision were calculated for different search approaches leveraging PharmGKB. Sixth, PharmGKB-enabled search strategies were compared to alternative strategies based on the use of MEDLINE. Finally, we analyzed the unfound manuscripts for their characteristics and importance. All analyses were conducted in February and March 2009.

Selection of Gold Standard Reviews. Two EGAPP reviews were selected as gold standard reviews: a 2007 systematic review evaluating the use of cytochrome P450 genotyping to guide the treatment of non-psychotic depression using SSRIs\(^8\) and a 2009 systematic review evaluating the use of UGT1A1 genotyping to guide the treatment of metastatic colorectal cancer using irinotecan.\(^9\) Furthermore, a 2008 rapid-ACCE review of CYP2C9 and VKORC1 genotyping to guide warfarin dosing\(^7\) was selected. These reviews were selected because they focused on PGx, used rigorous methods, and were associated with a substantial body of literature with which to conduct the analysis.

Abstraction and Classification of Included Manuscripts. To establish the search targets, the peer-reviewed manuscripts included in the systematic reviews were identified and classified as pertaining to analytic validity, clinical validity, and/or clinical utility. We did not include manuscripts if they were used solely to identify allele frequencies among sub-populations, as such estimates are directly provided in PharmGKB. Also, we considered excluding articles related to analytic validity, as genotyping platforms continue to improve rapidly, and as such validation can now be directly conducted using validated quality control materials, such as those available from the CDC Genetic Testing Quality Control Materials Program.\(^11\) However, we retained these manuscripts to be comprehensive in our analysis, while focusing our attention on the findings related to clinical validity and clinical utility.

Identification of PharmGKB Manuscripts. PharmGKB was searched for curated manuscripts describing the association between the drugs of interest and the genotypes of interest. Specifically, the manuscripts associated with the following drug-gene relationships were abstracted: irinotecan and UGT1A1; warfarin and (VKORC1 or CYP2C9); and (antidepressants or citalopram or fluoxetine or fluvoxamine or sertraline or escitalopram) and any of the cytochrome P450 genes. Each manuscript was manually classified as a review article or a non-review article. Only manuscripts published before the relevant gold standard reviews were included.

Citation Analysis. As the first step in the citation analysis, gold standard manuscripts were cross-referenced against PharmGKB manuscripts to look for direct matches. Next, we determined whether PharmGKB manuscripts cited the gold standard manuscripts in their reference lists. We believe this approach is justified, as most systematic reviews incorporate the systematic searching of reference lists within relevant manuscripts. To conduct this analysis, we used the Web of Science™\(^12\) to identify manuscripts that cited the gold standard manuscripts. Also, for six manuscripts not found in the Web of Science, Google™ Scholar\(^13\) was used to identify citing manuscripts. This citation network was used to determine whether a gold standard manuscript was cited by a PharmGKB manuscript.

Calculation of Recall and Precision. Recall and precision were calculated as follows: recall = gold standard manuscripts identified by search / all gold standard manuscripts, and precision = gold standard manuscripts identified / all manuscripts searched. These rates were calculated for the following...
strategies: (i) search PharmGKB manuscripts; (ii) search PharmGKB manuscripts and their reference lists; and (iii) search PharmGKB manuscripts and their reference lists, but only search the reference list if the manuscript is a review article. These rates were also calculated for alternative MEDLINE-enabled search strategies, as discussed next. We did not expand the denominator for precision calculations involving reference list searching, as searching for other relevant manuscripts within a reference list is a common practice within systematic reviews.

**Alternate Search Strategy.** The strategies just described were compared against a strategy involving the analysis of MEDLINE-identified review articles. In this strategy, relevant reviews were identified using the following search parameters: limit (review articles) & English language & published on or after 2000 but before the relevant gold standard review & test-specific query parameters. The test-specific query parameters were (i) serotonin uptake inhibitors (MeSH; explode) & cytochrome P-450 enzyme system (MeSH; explode); (ii) (UDP-glucuronosyltransferase 1A1 or UGT1A1) & irinotecan; and (iii) warfarin (MeSH) & [VKORC1 or cytochrome P-450 enzyme system (MeSH; explode)].

**Analysis of Unfound Manuscripts.** We assessed the characteristics and importance of manuscripts related to clinical validity and clinical utility that were not identified even by the search strategies with the highest recall rates. For this analysis, we assessed whether unfound manuscripts differed from found manuscripts with regard to sample size, study population, time between publication and systematic review, journal impact factor as determined using Journal Citation Reports, and citation frequency. Also, because of differences in search performance, we assessed whether SSRI manuscripts differed from manuscripts related to the other two topics. Using SAS® 9.1 (Cary, NC), the statistical significance of differences was evaluated using two-sample t-tests when variances were equal and Satterthwaite’s approximate t-tests when variances were unequal. All analyses were conducted at an alpha level of 0.05.

Finally, relevant domain experts assessed whether the omission of a manuscript, by itself and/or in combination with the other missing manuscripts, was likely to have changed the conclusions of the gold standard review. For warfarin, this analysis was conducted by an internist who has authored several peer-reviewed articles on the topic (DV). For SSRIs, this analysis was conducted by an internist who co-authored the gold standard review in question (LO).

**RESULTS**

**Identification of Relevant Manuscripts.** The results of the manuscript identification processes are summarized in Table 1. In brief, 77 gold standard manuscripts were identified across the three systematic reviews, 54 of which pertained to clinical validity and 1 of which pertained to clinical utility. PharmGKB included 114 manuscripts on the three topics, of which 20 were review articles. The MEDLINE search identified 86 reviews.

**Recall and Precision of Search Strategies.** The performance of the various search strategies evaluated is summarized in Table 2. Gold standard manuscripts for analytic validity had poor recall rates regardless of search strategy, ranging from 5% to 36%. Also, only 33% of manuscripts related to clinical validity or utility were directly curated in PharmGKB. However, recall was 78% and precision was 38% when the reference lists of PharmGKB articles were also searched (strategy C, Table 2).

The PharmGKB-enabled approach achieved a higher recall rate than the approach based on the use of MEDLINE to identify relevant reviews. In particular, of 23 gold standard manuscripts examining the clinical validity or utility of cytochrome P450 genotyping for guiding SSRI use, the MEDLINE-based strategy identified only 39% of the relevant gold standard manuscripts, whereas the PharmGKB-based approach identified 57%.

When used in conjunction, MEDLINE-identified reviews improved the performance of PharmGKB-based searches. Recall for articles related to clinical validity and utility was optimized by searching directly included PharmGKB manuscripts as well as the works cited by the reviews identified through PharmGKB and/or MEDLINE (strategy E, Table 2).

**Analysis of Unfound Manuscripts.** Based on the above findings, search strategies C and E in Table 2 appear to have the most favorable balance between recall and the effort required to conduct the search. However, neither strategy would be useful if
it missed many important manuscripts. Thus, we evaluated the importance and impact of manuscripts related to clinical validity that were not found by either of these search strategies. These manuscripts consisted of two articles related to warfarin PGx (references 36 and 52 in the original review) and ten manuscripts related to SSRI PGx (references 66, 68, 75-77, 79, 83, 87, 89, and 90 in the original review).

None of the missing studies randomized subjects to an intervention by genotype, but this was similar to the studies that were identified. Their conclusions were also similar to those that were identified.

As outlined in Table 3, unfound articles had significantly smaller sample sizes (i.e., fewer study subjects) than found manuscripts (p = 0.019). Consequently, the 78% of manuscripts identified by strategy C encompassed 87% of the 6869 subjects analyzed across all studies. Moreover, while published in a similar timeframe as the identified manuscripts, unfound manuscripts were published in journals with significantly lower impact factors (p = 0.004), and their citation rate per year was almost an order of magnitude lower (17.4 vs. 2.3, p < 0.0001).

Of note, unfound manuscripts were much more likely to involve Asian subjects (p = 0.023). As race is an important indicator of PGx allele frequencies, it is possible that Western investigators may have preferentially cited studies conducted in the West.

As a whole, SSRI manuscripts were similar to the unidentified manuscripts (Table 4). In particular, SSRI articles were 4.5 times less likely to be cited compared to non-SSRI articles (p < 0.0001). As our search strategies rely heavily on articles being cited by others, this fact likely explains why all search strategies had difficulty recapturing the SSRI studies.

Finally, as might be expected from their limited literature citation and smaller sample sizes, domain experts found that the omission of the 12 unidentified articles, either alone or as a group, was unlikely to have altered the conclusions of the original reviews.

**DISCUSSION**

**Summary and Interpretation of Findings.** Using three EGAPP and ACCE systematic reviews as gold standards, we investigated whether the PharmGKB research knowledge base could facilitate systematic assessments of PGx tests by assisting with
the efficient identification of peer-reviewed manuscripts relevant to clinical validity and utility.

Somewhat unexpectedly, we found that only 33% of such manuscripts were directly included in PharmGKB. This finding may be due to PharmGKB’s comprehensive scope, which may limit its depth of coverage for any single topic. We did find, however, that 78% of the gold standard manuscripts were recalled through the searching of PharmGKB-identified manuscripts and their reference lists. Precision was also reasonable, at 38%. The identified articles encompassed close to 90% of the study subjects, and these articles were cited 7.6 times more often than the unidentified articles. Also, domain experts determined that the omission of the remaining manuscripts was unlikely to have changed the conclusions of the original systematic reviews.

**Strengths.** As one strength, our search strategies would likely be faster and less costly to complete than traditional systematic review search strategies, while maintaining reasonable recall rates. Second, we used rigorous systematic reviews as gold standards. Third, we minimized bias due to the publication of the gold standard reviews by excluding manuscripts published after those reviews. Finally, we evaluated multiple search strategies that might be reasonably considered by clinicians, scientists, or policy makers.

**Limitations.** As one limitation, the publication of a gold standard review may have influenced the inclusion of manuscripts into PharmGKB. However, we excluded articles published after the systematic reviews, and the best search performance was achieved for the review published within months of this analysis. Second, we only included a limited number of PGx topics in the evaluation. Third, we did not quantitatively assess the time and resource requirements of our search strategies compared to traditional strategies. Fourth, because the precision of our search strategies is still suboptimal, the use of these strategies would need to be accompanied by a filtering process to identify truly relevant articles.

**Potential Search Strategy Based on Findings.** Based on our findings, we suggest the following search strategy when evaluating the clinical validity and utility of PGx testing. First, MEDLINE and the EGAPP Web site (http://www.egappreviews.org) should be screened to see if a definitive systematic review on the topic has already been published. If such definitive reviews are not available, a systematic review using a gold standard methodology such as EGAPP should be considered. However, if such an approach is impractical given available resources, and if it is acceptable to potentially miss some non-critical manuscripts, then PharmGKB manuscripts curated for the gene-drug relationships of interest should be examined. Also, the review should screen (i) works cited within all PharmGKB manuscripts or (ii) works cited within reviews identified through PharmGKB or MEDLINE. Finally, reviewers should consider (i) consulting domain experts and (ii) searching MEDLINE for relevant primary articles that may have been published too recently to have been cited by others or curated within PharmGKB.

**Implications and Future Directions.** We anticipate that our findings will facilitate the more efficient evaluation of the clinical validity and utility of PGx tests. Moving forward, we plan to continue to assess strategies for accelerating the translation of published PGx research findings into actionable information that can be used to guide clinical care.

**ACKNOWLEDGMENTS**

This work was supported by Award Number K01HG004645 from the National Human Genome Research Institute (KK). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health.

**REFERENCES**

14. Journal Citation Reports. Available from: http://www.thomsonreuters.com/products_services/scientific/Journal_Citation_Reports.