INTEGRATING VA'S NDF-RT DRUG TERMINOLOGY WITH PHARMGKB: PRELIMINARY RESULTS

JYOTISHMAN PATHAK, PhD
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: pathak.jyotishman@mayo.edu

LAURA C. WEISS*
Bethel University
3900 Bethel Drive, St. Paul, MN, USA
Email: laura-weiss@bethel.edu

MATTHEW J. DURSKI, BS
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: durski.matthew@mayo.edu

QIAN ZHU, PhD
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: zhu.qian@mayo.edu

ROBERT R. FREIMUTH, PhD
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: freimuth.robert@mayo.edu

CHRISTOPHER G. CHUTE, MD, DrPH
Department of Health Sciences Research, Mayo Clinic
200 1st Street SW, Rochester, MN, USA
Email: chute@mayo.edu

Abstract

Biomedical terminology and vocabulary standards play an important role in enabling consistent, comparable, and meaningful sharing of data within and across institutional boundaries, as well as ensuring semantic interoperability. The Veterans Affairs (VA) National Drug File Reference Terminology (NDF-RT) is a federally recommended standardized terminology resource encompassing medications, ingredients, and a hierarchy

* This work was done while the author was a summer intern at Mayo Clinic.

for high-level drug classes. In this study, we investigate the drug-disease relationships in NDF-RT and determine how PharmGKB can be leveraged to augment NDF-RT, and vice-versa. Our preliminary results indicate that with additional curation and analyses, information contained in both knowledge resources can be mutually integrated.

1. Introduction

Standardized biomedical terminologies play an important role in enabling consistent representation and interoperability of healthcare data and information systems. Within the realm of pharmaceutical drugs and medications, NDF-RT[2] created and maintained by the U.S. Department of Veterans Affairs is one of the publicly available federal medication terminologies. The goal of NDF-RT is to allow various clinical information systems using different drug nomenclatures to share and exchange medication data efficiently[3]. NDF-RT includes information about drugs and ingredients, provides a way to link and map standard clinical drug names to other drug terminologies (e.g., RxNorm), and is updated on a monthly schedule. Additionally, NDF-RT contains a multi-axial hierarchical knowledge structure that classifies ingredients and drug products based on their therapeutic intent, mechanism of action, pharmacokinetics and other aspects (see Section 2) It also provides several ways to create "meaningful groups" for the analysis of medication data. Several research efforts in the recent past, including our own prior work[4,5], have studied different aspects of NDF-RT or classification of medication data and their applications in information exchange[6], linkage[7] and querying[8].

However, to the best of our knowledge, there are no existing studies investigating and comparing drug-disease relationships in NDF-RT and Pharmacogenomics Knowledge Base (PharmGKB[9]), which is a comprehensive resource on pharmacogenes (i.e., genes involved in modulating the response to drugs), their variations, pharmacokinetics, pharmacodynamic pathways, and their effects on drug-related phenotypes. In particular, it is not known if drug-disease relationships that are manually curated in PharmGKB can be leveraged for augmenting the information contained in NDF-RT, and vice versa. Such a mutual integration of information from two related and publicly available knowledge resources could identify areas for additional curation, facilitate exchange of information as well as make such information sources more robust, increasing their utility for research and clinical applications.

To this end, in this paper we report our preliminary findings in comparing and mapping the drug-disease relationships in PharmGKB with those in NDF-RT. In particular, we studied the pharmacodynamics (PD) and pharmacokinetics (PK) relationships between drugs and diseases in PharmGKB, and compared them with the drug-disease relationships in NDF-RT. Our results indicate that while both sources contain related information, additional curation and analyses is required to enable mutual information integration.

2. Background

2.1. *Veterans Affairs National Drug File Reference Terminology*

The National Drug File Reference Terminology (NDF-RT)[2] is created and maintained by the Department of Veteran Affairs (VA). It includes information about drugs and ingredients, but also contains a multi-axial hierarchical knowledge structure that classifies various ingredients and drug

products. In particular, NDF-RT uses a description logic-based formal reference model that groups drug products into the high-level drug classes for Chemical Structure (e.g., Acetanilides), Mechanism of Action (e.g., Prostaglandin Receptor Antagonists), Physiological Effect (e.g., Decreased Prostaglandin Production), drug-disease relationship describing the Therapeutic Intent (e.g., Pain), Pharmacokinetics describing the mechanisms of absorption and distribution of an administered drug within a body (e.g., Hepatic Metabolism), and legacy VA-NDF classes for Pharmaceutical Preparations (VHA Drug Class; e.g., Non-Opioid Analgesic). Figure 1 shows NDF-RT's content model.

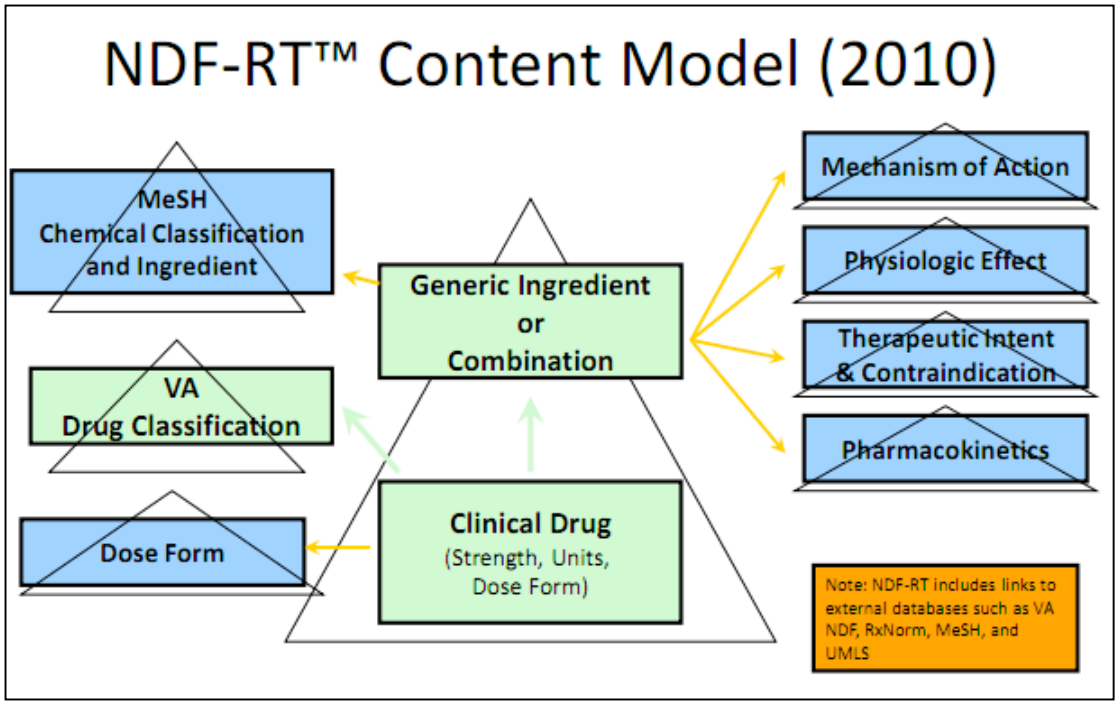

Figure 1 NDF-RT content model (Source: http://evs.nci.nih.gov/ftp1/NDF-RT)

This figure shows the structure of NDF-RT: triangles denote hierarchies of related concepts, each hierarchy being categorized in the rectangles within the triangles. Taxonomic or ISA relationships (upward-pointing green arrows) unify NDF-RT clinical drug concepts into a polyhierarchy, classified both by their VA drug class and their generic ingredient(s). Various named role relationships (sideways-pointing amber arrows) define the central drug concepts (green) from which they originate in terms of the reference hierarchy concepts (blue) pointed to. Role relationships are also inherited into subsumed clinical drug concepts. Note that NDF-RT also augments a "legacy" classification system (called VA-NDF[1]) which classified drug products into groupings developed by VA (denoted by VHA Drug Class) to support organization and decision support for medication usage in clinical care settings.

2.2. *The Pharmacogenomics Knowledge Base*

The Pharmacogenomics Knowledge Base (PharmGKB)[9] is a comprehensive resource for pharmacogenes (i.e., genes involved in modulating the response to drugs), their variations, pharmacokinetics, pharmacodynamic pathways, and their effects on drug-related phenotypes. Its overarching goal is to support the integration, aggregation and curation of the information contained in various life sciences and biological databases that contain information about genetic variation and associated phenotypes. The data in PharmGKB is curated from the literature and other databases that report genetic variations in known pharmacogenes. To facilitate indexing and retrieval of the information, the PharmGKB data sets are annotated with genes and/or drugs based on five different categories: clinical outcomes, pharmacodynamics and drug response, pharmacokinetics, molecular and cellular functional assays, and genotype (see Figure 2).

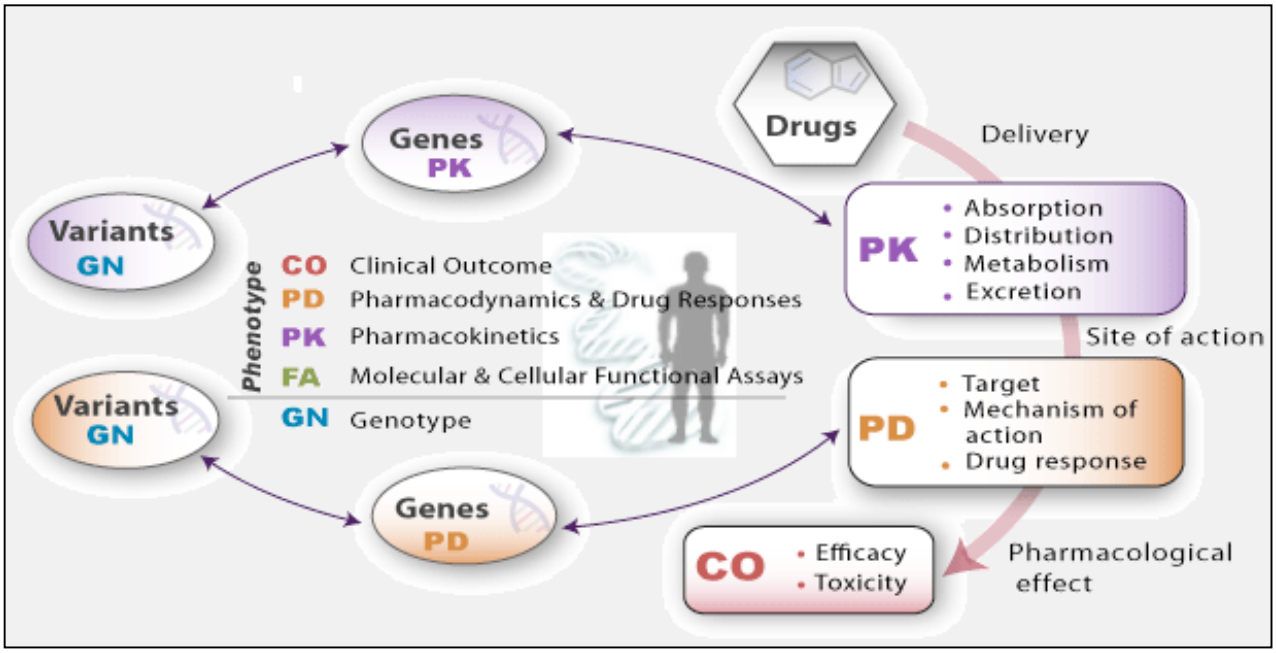

Figure 2 Pharmacogenomics information flow (Source: http://www.pharmgkb.org)

3. Materials and Methods

3.1. *Materials*

The primary materials used in this study are the following:

- The July 6, 2011 PharmGKB relationships data set, available for download via http://www.pharmgkb.org/resources/downloads_and_web_services.jsp. The full data set contains a total of 24,213 relationships between the genes, drugs and diseases in PharmGKB (see Table 1). For this study, our investigative efforts focused exclusively on the 2,697 drug-disease relationships in PharmGKB. This data set included all four types of drug-disease relationships: both PD and PK, PD only, PK only, and neither PD nor PK (see Table 1). For example, the drug *Imatinib* has several types of relationships with different diseases: both PD and PK with the disease *Gastrointestinal Stromal Tumors*, only PD with

the disease *Glioma*, only PK with the disease *Neoplasms*, and neither PD nor PK with the disease *Leukemia*.

	Type of Relationship						
	Drug-Drug*	Disease-Disease*	Gene-Gene*	Drug-Disease	Gene-Disease	Gene-Drug	Total
PD Only	56	0	0	1,786	102	2,989	4,933
PK Only	230	0	0	137	1	1,280	1,648
Both PD and PK	47	0	0	446	5	986	1,484
No PD and No PK	303	145	735	328	8,166	6,471	16,148
Total	636	145	735	2,697	8,274	11,726	24,213

Table 1 Relationships in PharmGKB, before analysis (highlighted cells indicate relationships studied in this work)
*The numbers for these relationships include duplicates.

- The June 2011 NDF-RT ontology of drugs and diseases, accessible via http://evs.nci.nih.gov/ftp1/NDF-RT/. NDF-RT is organized as a relationship-based terminology source of medications and related entities (i.e. drugs, diseases, ingredients, etc.). Each unique term can be described via its relationship with some other term in the ontology. This identification process is further illustrated in NDF-RT's content model (see Figure 1), which is a visual representation of an individual term's organizational structure. For this study, we investigated the information contained in the *Role Relationships* section of a particular drug. This feature of NDF-RT connects a specific drug with a variety of other factors, such as an ingredient, mechanism of action, or disease (see Table 2). Similar

Type of Relationship	Total
has_ingredient	23,244
has_MoA (Mechanism of Action)	14,749
has_PE (Physiologic Effect)	25,370
has_PK (Pharmacokinetic)	750
has_TC (Therapeutic Category)	76
may_diagnose	954
may_treat	47,636
may_prevent	5,968
Total	118,747

Table 2 Relationships in NDF-RT, before analysis (highlighted cells indicate relationships studied in this work)

to the restrictions for PharmGKB, in this study we were concerned only with the relationships between a drug and disease. Therefore, we limited our scope to include four relationship types from NDF-RT: 'has_PK', 'may_treat', 'may_prevent', and 'may_diagnose'. It should be noted that while NDF-RT contains a 'has_PE' relationship, due to ambiguity in the NDF-RT documentation about its relevance to the pharmacodynamics of a given drug, it was excluded from our evaluation.

3.2. *Methods*

For this study, we investigated the 2,697 drug-disease relationships from PharmGKB in comparison with NDF-RT. Of these 2,697 relationships in PharmGKB, we removed all relationships between drug classes and diseases (n=363) from analysis for uniformity with NDF-RT. The reason for doing this is because in order for a drug-disease relationship to exist in NDF-RT, the term must be an individual drug within a drug class, rather than the drug class itself. In other words, NDF-RT does not provide relationships between drug classes and diseases, but rather between individual drugs (that belong to a drug class) and diseases. Hence, our comparison process began with the remaining 2,334 drug-disease relationships from PharmGKB.

For each unique drug-disease pair in the PharmGKB data set, we searched for an identical drug-disease pair in NDF-RT. If an appropriate match was identified, we recorded the match and type of relationship in the data set. If no match was identified in NDF-RT, the match was designated as "missing". For example, the drug *Fluorouacil* has a PD relationship with *Colonic Neoplasms* in PharmGKB. A relationship between this drug and disease was identified in NDF-RT as a "may_treat" relationship. This identified match and its corresponding classification were recorded in the data set. Throughout our comparison process, we discovered several additional drug-disease relationships that existed in NDF-RT, but not in PharmGKB. The type of relationships for these pairs was also recorded and they were added to the data set. To continue with the example above, *Fluorouracil* had a 'may_treat' relationship with *Keratosis* in NDF-RT. An identical relationship of any type did not exist in PharmGKB; thus an entry for the relationship in NDF-RT was added to the data set for analysis.

This process was repeated for all 2,334 drug-disease relationship pairs in PharmGKB, adding entries from NDF-RT when necessary. The drug ID and disease ID was recorded for each pair in the list from each knowledge base, as well as the type of relationship in both PharmGKB (PD, PK, both, or neither) and NDF-RT (has_PK, may_treat, may_prevent, may_diagnose, or a combination of the four). Several of the drug-disease pairs in our final analysis data set have multiple types of relationships for the same pair. In PharmGKB, many pairs have both a PD and a PK relationship (e.g., drug *Adalimumab* and disease *Arthritis, Rheumatoid*). Additionally, there were a large number of entries from NDF-RT that had both a 'may_treat' and 'may_prevent' relationship (e.g., drug *Aspirin* and disease *Pain*).

Throughout our comparison, there were a total of 57 entries in which the drug was matched with a synonym for the drug. For example, PharmGKB's *Salbutamol* was identified in NDF-RT as

Albuterol, a synonym for *Salbutamol*. Similarly, there were 22 diseases that were matched to a synonymous disease name. Furthermore, 77 drug entries from PharmGKB were matched with a drug in NDF-RT that did not have a *Role Relationships* section available (e.g., *Nitrazepam*). For these, the PharmGKB drug-disease pairs were recorded as not having a match in NDF-RT and no additional entries were added to the data set from NDF-RT. Finally, the drugs in 168 entries from PharmGKB did not have any appropriate match in NDF-RT and could not, therefore, be matched with a suitable drug in NDF-RT (e.g., *Orbofiban*). These outlying cases were all included in the final analysis, as they have potential to augment and clarify both PharmGKB and NDF-RT.

4. Results

The comparison between PharmGKB and NDF-RT began with the 2,334 drug-disease relationships in PharmGKB. Of these 2,334 relationships, 2,039 existed only in PharmGKB, while 295 were common to both PharmGKB and NDF-RT. By the end of the comparison, a total of 1,499 relationships identified in NDF-RT were added to the original data set for analysis. These additional entries were drug-disease pairs that existed only in NDF-RT, and not in any form in PharmGKB. See Table 3 for a summary of these results.

Relationship	Number of Relationships
Common to both PharmGKB and NDF-RT	295
Exist only in PharmGKB	2,039
Total in PharmGKB (starting number for analysis)	2,334
Exist only in NDF-RT (added to data set for analysis)	1,499
Total from PharmGKB and NDF-RT for analysis	3,833

Table 3 Relationships from PharmGKB and NDF-RT studied in this work (highlighted cells indicate relationships studied in this work)

In order to augment the relationships in PharmGKB using NDF-RT, and vice versa, each drug-disease relationship must be classified according to its type. The only relationship type that is common to both PharmGKB and NDF-RT in their current releases is PK. Therefore, throughout our investigation, we were primarily focused on any PK relationship from one knowledge base that could be augmented in the other. For example, PharmGKB's PK relationship between the drug *Cocaine* and the disease *Apnea* could potentially be added to NDF-RT. As demonstrated in Table 4, there are a total of 537 PK relationships in PharmGKB that are currently not represented in NDF-RT under the relationship type has_PK, although they may be classified under some other relationship type (e.g., 'may_prevent'). For example, PharmGKB has an identified PK relationship between the drug *Lozartan* and the disease *Hypertension*. While this relationship is captured in NDF-RT, the drug-disease pair is represented as a may_treat relation, instead of

has_PK. As for the has_PK relationships identified in NDF-RT, all 34 do not exist in PharmGKB under any relationship type.

	Number of relationships, per classification				
	PK	PD	may_treat	may_prevent	may_diagnose
Total in PharmGKB (not represented in NDF-RT)	537 (537)	1,942 (1,942)	N/A	N/A	N/A
Total in NDF-RT (not represented in PharmGKB)	34 (34)	N/A	1,662 (1,662)	183 (183)	8 (8)

Table 4 Comparison between drug-disease relationships in PharmGKB and NDF-RT, according to classification

Unlike the PK relationship, PharmGKB's PD classification does not currently exist in NDF-RT. Therefore, all PD relationships in PharmGKB (1,942) are unique to that data set and could not be added to NDF-RT for evaluation. The same can be said for NDF-RT's 'may_treat', 'may_prevent', and 'may_diagnose' relationships. As these classifications do not currently exist in PharmGKB, a relationship of the same type could not be identified. For example, NDF-RT has a 'may_prevent' relationship between the drug *Flunisolide* and the disease *Asthma*. Although this relationship does in fact exist in PharmGKB with a relationship type of PD, a 'may_prevent' classification is not yet available. Therefore, we conclude that all relationships of these types in NDF-RT do not exist under the same classification in PharmGKB.

Due to PharmGKB not having the 'may_treat', 'may_prevent', and 'may_diagnose' relationships in its current release, it is not possible to compare these types of relationships in PharmGKB and therefore they have been marked as "not applicable" in Table 4. Similarly, there were no PD relationships identified in NDF-RT due to the absence of this relationship in NDF-RT's current release. See Table 4 for a summary of these results.

5. Discussion

The principal goal of this study was to identify and analyze the relationships in NDF-RT that could be used to augment PharmGKB, and vice versa. As the PK relationship is common to both knowledge bases, this relationship is the most immediately applicable finding for our purposes. A total of 537 PK relationships were identified in PharmGKB and 34 has_PK relationships were identified in NDF-RT. Interestingly, of these 571 drug-disease PK pairs, none were common to both knowledge bases. One plausible reason for this is the lack of a consistent definition of 'has_PK', 'may_treat', and 'may_prevent' relationships in NDF-RT with the PK relation in PharmGKB. Another reason for non-overlap might be due to the focus and scope of these knowledge resources: NDF-RT is primarily developed and maintained for encoding clinical data, whereas the main focus of PharmGKB is for biological and life sciences. However, in spite of these issues, these results indicate that there is a potential for unification and curation of the PK relationships in the two knowledge sources. Such an effort should encourage the development of a common set of gene-drug, gene-disease, and drug-disease relationships so that information is more easily comparable and integrable.

The comparative process described in this work began with the drug-disease pairs in PharmGKB that had both a PD and PK relationship. Throughout this work, we emphasized the fact that our study originated with these specific entities, and expanded to include additional relationships after comparison with NDF-RT. Some of these additional relationships that were added from NDF-RT were found to be present in PharmGKB, though without *both* a PD and PK relationship. For this reason, a number of relationships that have just a PD, or just a PK, or neither in PharmGKB have been included in our analysis.

From this study, we also determined that on several occasions the relationship type (i.e., PD or PK) is not specified between a drug-disease pair in PharmGKB. While we acknowledge that this under-specification of the relationship type could be due to limited scope of curation (i.e., only selected journals are curated by PharmGKB curators), not having a proper designation of the type of relationship between the drug and disease entities that exists in the PharmGKB can lead to confusion. This reiterates the goal of this study to investigate publicly available knowledge resources, such as NDF-RT, for further evaluation and mutual information integration. For example, it might be worth investigating the incorporation of NDF-RT relationship types, such as 'may_treat', 'has_MOA' etc., within the PharmGKB knowledge base. The same could be said about NDF-RT's potential to include a PD relationship. Furthermore, one might consider investigating Structured Product Labels to perform a similar analysis between drugs and adverse drug reactions.

Finally, this study focused exclusively on drug-disease relationships, as they are common to both knowledge bases. However, PharmGKB encompasses gene-disease, drug-drug and gene-drug relationships, as well. Integrating these relationships in NDF-RT could be beneficial in addition to the drug-disease relationships that currently exist in NDF-RT. We plan to pursue an extended analysis encompassing such relationships in the future.

6. Conclusion

In this study, we investigate the drug-disease relationships in two publicly available knowledge resources, namely PharmGKB and NDF-RT. Our results indicate that with additional curation and authoring of relationships, both resources can be mutually integrated.

7. Acknowledgments

This work is supported by the PGRN Pharmacogenomic Ontology Network Resource (PHONT; U01GM061388).

References

1. Nelson SJ, Brown SH, Erlbaum MS, et al. A Semantic Normal Form for Clinical Drugs in the UMLS: Early Experiences with the VANDF. *AMIA Annual Symposium.* 2002:557-561.

2. Brown S, Elkin P, Rosenbloom T, et al. VA National Drug File Reference Terminology: A Cross-Institutional Content Coverage Study. *MedInfo: Studies in Health Technology and Informatics.* 2004:477-781.

3. Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. In: Geissbuhler A aKC, ed. *IMIA Yearbook of Medical Informatics.* Vol 47: International Medical Informatics Association; 2008:67-79.

4. Pathak J, Chute C. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. *Journal of American Medical Informatics Association.* 2010;17(4):432-439.

5. Pathak J, Chute C. Further revamping VA's NDF-RT drug terminology for clinical research. *Journal of American Medical Informatics Association.* 2011;18(3):347-348.

6. Bouhaddou O, Warnekar P, Parrish F, et al. Exchange of Computable Patient Data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Mediation Strategy. *Journal of the American Medical Informatics Association.* 2008;15(2):174-183.

7. Burton MM, Simonaitis L, Schadow G. Medication and Indication Linkage: A Practical Therapy for the Problem List? *AMIA Annual Symposium.* 2008:86-90.

8. Palchuk M, Klumpennar M, Jatkar T, Zottola R, Adams W, Abend A. Enabling Hierarchical View of RxNorm with NDF-RT Drug Classes. *AMIA Annual Symposium*. Washington, DC2010:577-581.

9. Altman RB. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet.* 2007;39(4):426-426.