Review

Semantically enabling pharmacogenomic data for the realization of personalized medicine

Matthias Samwald
Section for Medical Expert and Knowledge-Based Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna
Spitalgasse 23, A-1090 Vienna, Austria
&
Institute of Software Technology and Interactive Systems, University of Technology Vienna
Favoritenstrasse 9-11/188, A-1040 Vienna, Austria
Tel: +43 140400 6665
matthias.samwald@meduniwien.ac.at

Adrien Coulet
LORIA – INRIA Nancy - Grand-Est,
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex, France
Tel: +33 3 54 95 86 38
adrien.coulet@loria.fr

Iker Huerga
Linkatu, S.L.
Polo de Innovación Garaia. Goiru 1, Edificio A, 4º Piso.
Mondragón - Guipuzkoa, 20500 Spain
Tel: +34 943 712 072
ihuerga@linkatu.net

Robert L. Powers
Predictive Medicine, Inc.
37 Mansion Drive,
Topsfield, MA 01983, USA
Tel: +1 978 887 9150
rpowers@predmed.com

Joanne S. Luciano
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8th Street, Winslow 2143
Troy, NY 12180, USA
Tel. +1 518 276 4939

jluciano@cs.rpi.edu

Robert R. Freimuth
Division of Biomedical Statistics and Informatics, Department of Health Sciences Research
Mayo Clinic
200 First Street SW
Rochester, MN 55901
Tel: +1 507 284 5541
freimuth.robert@mayo.edu

Frederick Whipple
Genomics Education Initiative
609 Sycamore Avenue
Fullerton, CA 92831
Tel: +1 714 616-2669
fwhipple@SNPsAndChips.com

Elgar Pichler
Department of Chemistry and Chemical Biology
Northeastern University
360 Huntington Avenue
Boston, MA 02115
Tel: +1 781.431.0359
elgar.pichler@gmail.com

Eric Prud'hommeaux
World Wide Web Consortium / MIT
32 Vassar Street,
Cambridge, MA 02140, USA
Tel: +1 617 258 5741
eric@w3.com

Michel Dumontier
Department of Biology, School of Computer Science, Institute of Biochemistry
Carleton University
1125 Colonel By Drive
Ottawa, ON K1T2T4
Tel: +1 613 520 2600 x4194
michel_dumontier@carleton.ca

M. Scott Marshall
Department of Medical Statistics and Bioinformatics, Leiden University Medical Center / Informatics Institute, University of Amsterdam
Einthovenweg 20, 2333 ZC Leiden
Tel: +31 71 5269111
marshall@science.uva.nl

Summary

Understanding how each individual's genetics and physiology influences the pharmaceutical response is crucial to the realization of personalized medicine and the discovery and validation of pharmacogenomic biomarkers is key to its success. However, integration of genotype and phenotype knowledge in medical information systems remains a key challenge. The inability to easily and accurately integrate the results of biomolecular studies with patients' medical records and clinical reports prevents us from realizing the full potential of pharmacogenomic knowledge for both drug development and clinical practice. Herein, we describe approaches using Semantic Web technologies, in which pharmacogenomic knowledge relevant to drug development and medical decision support is represented in such a way that it can be efficiently accessed both by software and human experts. We suggest that this approach increases the utility of data, and that such computational technologies will become an essential part of personalized medicine, alongside diagnostics and pharmaceutical products.

Keywords: *ontologies, pharmacogenomics, translational medicine, personalized medicine, clinical decision support systems, knowledge representation*

Future perspective

Personalized medicine involves the customization of therapies based on genetic, environmental and physiological factors to deliver the best possible care to individual patients. In the past decade, there has been a substantial shift in our understanding of factors that influence therapeutic outcomes. This has been driven largely by the increasing availability of large scale profiling technologies, including next generation sequencing and gene expression profiling. One of the prominent applications of molecular profiles in clinical practice is to stratify the patient population in order to identify positive, neutral and negative responders. Widespread use of profiling technologies has become significantly more feasible for inclusion in clinical trials as costs have been decreasing [1]. Indeed, the ability to build and test patient profiles against therapeutic outcomes in clinical trials may increase the number of approved therapies overall, with the resulting drugs and therapies approved for clearly specified segments of the population. Additional benefits for pharmaceutical companies include the opportunity to develop incrementally modified drugs with reduced risk for drug development [2]. While all of these developments promise to change the practice of medicine, a remaining challenge lies in the effective integration of highly heterogeneous data obtained from a variety of sources, and its use in accurate predictive systems for supporting knowledge discovery and personalized clinical decision making.

The role of pharmacogenomics in tailored therapies

In current practice a drug dose is adjusted according to generic factors such as weight, age and kidney and liver function. However, two individuals with the same values for these parameters may respond differently to the same therapy, due primarily to differences at the genetic level that affect either the rate at which the drug is metabolized or the susceptibility of the target to

modulation. Pharmacogenomic studies attempt to link genetic variation with differences in the drug responses. The use of an individual's genetic information to select drugs and specify dosages lies at the heart of personalized drug-based therapies. For example, warfarin is one of over a dozen drugs approved by the U.S. Food and Drug markers are provided (table 1). In the US alone, over 20 million prescriptions of warfarin are written per year to treat chronic anti-coagulation indications including atrial fibrillation, deep vein thrombosis, pulmonary embolism and artificial heart valves. Over- or underdosing can have serious consequences: warfarin is one of the leading causes of emergency care and drug-related hospitalization due to adverse drug events [3]. Genetic variations in two key genes have been found to strongly affect the toxicity of warfarin and consequently its initial recommended dose. First, the cytochrome P450 2C9 enzyme (CYP2C9) influences the overall amounts of warfarin in the bloodstream. Second, warfarin's activity depends on the variant of the gene encoding a subunit of its drug target, the enzyme *vitamin K epoxide reductase*.

Table 1: Range of Expected Therapeutic Warfarin Doses (mg/day) based on CYP2C9 and VKORC1 genotypes as described in an FDA drug label. Reproduced from the updated warfarin (Coumadin, Bristol-Myers Squibb, Princeton, New Jersey) product label. Dosage recommenations in grey deviate from standard dosage recommenations because of pharmacogenetic findings. CYP2C9 = cytochrome P450 2C9; FDA = U.S. Food and Drug Administration; VKORC1 = vitamin K epoxide reductase complex, subunit 1.

genotype		*CYP2C9*					
		**1/*1*	**1/*2*	**1/*3*	**2/*2*	**2/*3*	**3/*3*
VKORC1	*GG*	5-7 mg	5-7 mg	3-4 mg	3-4 mg	3-4 mg	0.5-2 mg
	AG	5-7 mg	3-4 mg	3-4 mg	3-4 mg	0.5-2 mg	0.5-2 mg
	AA	3-4 mg	3-4 mg	0.5-2 mg	0.5-2 mg	0.5-2 mg	0.5-2 mg

While the warfarin dosage table is relatively simple to interpret, the inclusion of additional factors and more complex dosing algorithms, which have been demonstrated to improve warfarin dosage [4–6], requires a more sophisticated, computer-assisted system for individualizing therapies. Furthermore, it can be expected that clinically relevant pharmacogenomic findings will be discovered for a rapidly growing number of drugs, so that pharmacogenomic considerations will be relevant for not only a few, special therapies, but for a large fraction of commonly prescribed drugs.

As the findings and guidelines for optimizing pharmacotherapy become more complex and more widespread, they are likely to become so overwhelming for clinicians that they are not applied consistently in daily practice. If this occurs, the benefits of treatment optimization based on pharmacogenomic knowledge will not reach patients. To realize the full potential of pharmacogenomics, the use of computer-based decision support systems will become indispensable. The move to more advanced decision support in clinical practice would also make it possible to consider drug-drug interactions that influence drug or target activity.

Semantic Infrastructure for the Biomedical Sciences

In this paper we report on progress toward the development of a global computational infrastructure for pharmacogenomics in the context of personalized medicine. At the heart of

this infrastructure are the so-called Semantic Web technologies produced by the World Wide Web Consortium (W3C). These technologies aim to facilitate the representation and processing of datasets containing increasingly sophisticated knowledge. Semantic Web technologies are being adopted worldwide by organizations that want to leverage technology built for the World Wide Web in order to publish, share, query and integrate data with others. Hundreds of datasets have been linked in this way, resulting in a global cloud of interlinked data. We will identify datasets and vocabularies of interest for pharmacogenomics from which decision support systems may be developed in the future.

Semantic Web technologies are based on two ideas: resolvable identifiers and machine understandable descriptions. Uniform Resource Identifiers (URI) can be used to identify any entity, whether it is a hospital, a dosage regime, a kind of drug, a kind of genetic variation, or even a clinical report. An example of a URI is that for warfarin (http://bio2rdf.org/drugbank_drugs:DB00682) as defined by Drugbank database, but which is provided by the Bio2RDF project [7]. Entities identified by URIs can be described in terms of their attributes and the relations they hold with other entities. The Resource Description Framework (*RD*F, [8]) provides a simple model in which statements are captured using subject-predicate-object *triples*, where the predicate indicates a relation between the subject and the object. So, a statement that warfarin is the ingredient of a drug with the brand name Coumadin would be written as:

```
<http://bio2rdf.org/drugbank_drugs:DB00682>
  < http://bio2rdf.org/drugbank_ontology:brandName>
    "Coumadin"
```

A statement that warfarin (as defined by DrugBank entry) is related to warfarin (DB00682as defined by ChEBI entry 10033) would be written:

```
<http://bio2rdf.org/drugbank_drugs:DB00682>
 <http://bio2rdf.org/bio2rdf_resource:xRef>
    <http://bio2rdf.org/chebi:10033>
```

Based on such triples, complex networks of interlinked statements can be built. This makes it possible to navigate and aggregate globally distributed data, enabling the transparency and scalability that made the Word Wide Web one the most successful technologies in recent history.

Slightly more sophisticated than RDF is OWL, the Web Ontology Language [9]. OWL is based on formal logics and can be used to capture general rules about the world (such as “every person has two two biological parents”, “every CYP2C9*2 allele has a thymidine nucleoside at position 430”). It has been used on many occasions to formally represent pharmacogenomics knowledge so that it becomes possible to answer questions that require automated reasoning [10,11].

Linking Open Data

Linked Open Data (LOD) is a set of design principles which make it practical to share machine-readable information on the web. The set of services following these principles has come to be called the Linked Open Data cloud. This cloud is growing exponentially, forming a large, distributed data store collecting cultural, geographic, political, economic, scientific and other information. The distributed nature of these databases enables independent contributions, which is critical to growing knowledge at the scale required to capture the domains intrinsic to solving health care and pharmacology problems.

At the core of the LOD cloud is RDF's use of IRIs (think URLs) for distributed extensibility. Using IRIs to represent e.g. chemicals and the relationships between them, encourages others to use the same identifiers, and enables consumers to be confident about what concepts the data publisher is trying to convey. Identifiers from Uniprot and KEGG are widely used in the around 25 biological LOD data sets.

LOD is designed to encourage the expression of linkages between data. The network basis for RDF allows us to express the interconnectedness of databases like DailyMed, DrugBank and SIDER, and to ask questions which reflect the real complexities of our domain. A wide variety of pharmaceutical datasets have been made openly available in Semantic Web formats by the *Linked Open Drug Data* task force of the World Wide Web Consortium [12].

Information resources for pharmacogenomics

A number of public domain databases are available that can provide key information relevant for understanding genetic variation and its potential impact on disease and treatment. Of special interest are those curated databases that describe associations between genotypes and phenotypes in humans (some examples are provided in table 2*)*.

Table 2: Databases containing associations between genetic variations, associated phenotypes and genetic tests.

Pharmacogenomics Knowledgebase (PharmGKB)	A large database of curated knowledge and raw data about associations between genes, genetic variants, drug response and disease [13,14].
GWAS Central (formerly called HGVbaseG2P)	A database of genome-wide association studies that also provides summaries of study results [15].
SNPedia	A wiki-based platform containing information on phenotypes associated with SNP variants, population prevalence of genetic variants and SNP microarrays [16].
Online Mendelian Inheritance in Man (OMIM)	Information about diseases with Mendelian inheritance, including references to the implicated genes [17].
dbGaP	Results of studies that have investigated the interaction of

	genotype and phenotype [18].
GEN2PHEN Knowledge Center	Integrated genotype-to-phenotye data with facilities for data annotation and user feedback [19,20].
GET-Evidence	A large database of automatically annotated and then manually curated information about the impact of genetic variations [21].
HuGE Navigator	Information on genetic variants, gene-disease associations, gene-gene and gene-environment interactions, and evaluation of genetic tests [22].
Genetic Association Database (GAD)	Diseases associated with genetic variants [23].
Genotator	An aggregated gene-disease relationship data containing an integrated view over other datasets [24].
NCBI GeneTests	This resource concerns genetic tests used in diagnostic and genetic counseling [25].
The Genetic Testing Registry	A database (under development) about genetic markers and tests that enable their clinical exploration [26].

To create a comprehensive knowledge base about drugs and their pharmacogenomic properties, these data need to be combined with data about approved pharmaceuticals, ongoing clinical trials, drug interactions, clinical guidelines and other kinds of biomedical knowledge including observations that link genotypes to phenotypes. However, the scalable and sustainable integration of data from such a variety of large, complex, distributed and heterogeneous sources has proven to be a challenge. Over the past five years, the development of ontologies to integrate data has emerged as a key technology for addressing the challenge of data integration.

Ontologies and terminologies play a critical role in data integration. They enable users to use well-defined, unambiguous terms to semantically annotate their data, thereby providing the means by which one can query across different datasets which use the same terms. Terminologies and coding systems focus on providing a comprehensive set of terms. By contrast, ontologies are a formal representation for specifying the entities and attributes in a domain of discourse such as pharmacogenomics. When an ontology is expressed in OWL, automatic reasoning can be performed with a variety of free open source tools. Table 3 lists relevant ontologies and terminologies of interest to pharmacogenomics.

Table 3: Ontologies and terminologies of relevance for pharmacogenomics.

Types of represented information	**Name**	**Description**

All of translational and personalized medicine	Translational Medicine Ontology (TMO)	An ontology covering key aspects of the entire spectrum of translational and personalized medicine, developed by participants of the W3C Heath Care and Life Science Interest Group [27].
PGx	Suggested Ontology for Pharmacogenomics (SO-Pharm)	A complex ontology that represents phenotype, genotype, treatment and their relationships in groups of patients. SO-Pharm has been designed to guide knowledge discovery in pharmacogenomics [28].
PGx	Pharmacogenomics Ontology (PO)	An ontology built from PharmGKB, and includes biomedical measures and outcomes [29].
Mutation Impact	Mutation Impact ontology	An ontology designed to represent mutation impacts on protein properties resulting from an information extraction process [30].
Genotype	Sequence Ontology (SO)	Contains terms often used for the annotation of sequences and features, including detailed description of different types of sequence variations [31,32].
Chemical	ChEBI	Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds [33,34]
Chemical	RxNorm	Normalized names for clinical drugs, references to other terminologies [35,36].
Chemical, clinical	Logical Observation Identifiers Names and Codes (LOINC)	An established coding system for clinical lab results. Contains many identifiers for results of genetic [37,38].
Phenotype	Disease Ontology	An ontology of human diseases [39,40].
Phenotype	Phenotypic Quality Ontology (PATO)	An general ontology of qualities that can be used to describe phenotypes [41].
Phenotype	Human Phenotype Ontology	An ontology for phenotypic abnormalities encountered in human disease [42].
Anatomy	Foundational Model of Anatomy (FMA)	An ontology for the canonical, anatomical structure of an organism [43].
Safety	Medical Dictionary for Regulatory Activities (MedDRA)	A terminology currently for safety reporting (mandated in Europe and Japan for safety reporting, standard for adverse event reporting in the U.S.) [44].

The coverage of genetic information in established clinical coding schemes and ontologies varies. For example, *Logical Observation Identifiers Names and Codes (LOINC)* is an established standard for representing clinical laboratory results. The current version contains many identifiers for the results of genetic tests. On the other hand, SNOMED CT, one of the most widely employed general clinical terminologies, contains very few specific terms from the domains of genetics and pharmacogenomics. RDF/OWL makes it possible to merge these different coding schemes and ontologies in order to compile a comprehensive model covering all aspects of pharmacogenomics and its clinical context.

Extracting knowledge from the pharmacogenomics literature

A substantial amount of pharmacogenomic knowledge is captured in the scientific literature. Unfortunately, this knowledge is expressed in natural language, and is therefore difficult to integrate and use in combination with other structured data resources. To complicate matters, the volume of literature containing facts relevant to pharmacogenomics is large and continuously increasing.

To effectively extract pharmacogenomic facts from the literature, automated methods must be employed. Information extraction techniques such as natural language processing (NLP) as well as statistical models from machine learning can be used to identify entities of pharmacogenomic interest (such as genes, gene variants, drugs and drug responses) and the relations between these entities in unstructured text [45]. After extraction, entities and relations can be normalized with standard dictionaries and ontologies [46,47], and encoded in a structured format. Such normalized relations can subsequently be compared to other literature derived relations and to the content of other databases [48]. Furthermore, Semantic Web representations of the extracted normalized relations can be made available to a broader community of researchers, drug developers and medical practitioners.

At present, text collections used for text mining in biomedical research are mostly comprised of Medline abstracts. In the future, information extraction will increasingly be used on full text article collections and will allow a more complete extraction of gene-drug-phenotype relationships from scientific publications, clinical records, or the patent literature.

The use of information extraction techniques in pharmacogenomics resulted, for example, in the creation of tools for the extraction of pharmacogenomic concepts and relationships [49], the automatic construction of databases, in particular a side effect resource (SIDER) [50], the completion of pharmacokinetics pathways [51], the creation of a Pharmacogenomic Relationships Ontology (PHARE) [52], and the extraction of mutation impacts on protein properties which were used to populate the Mutation Impact Ontology [53].

Using semantically enabled pharmacogenomics data for personalized clinical decision support

Once all the different components described above are in place and have matured, a powerful system for the creation of decision support systems for personalized pharmacotherapy emerges (Fig. 1). While clinical decision support systems in the past often suffered from a lack of publicly

available, formally represented biomedical knowledge, the translation of the growing wealth of pharmacogenomic findings into rules for clinical decision support is relatively simple.

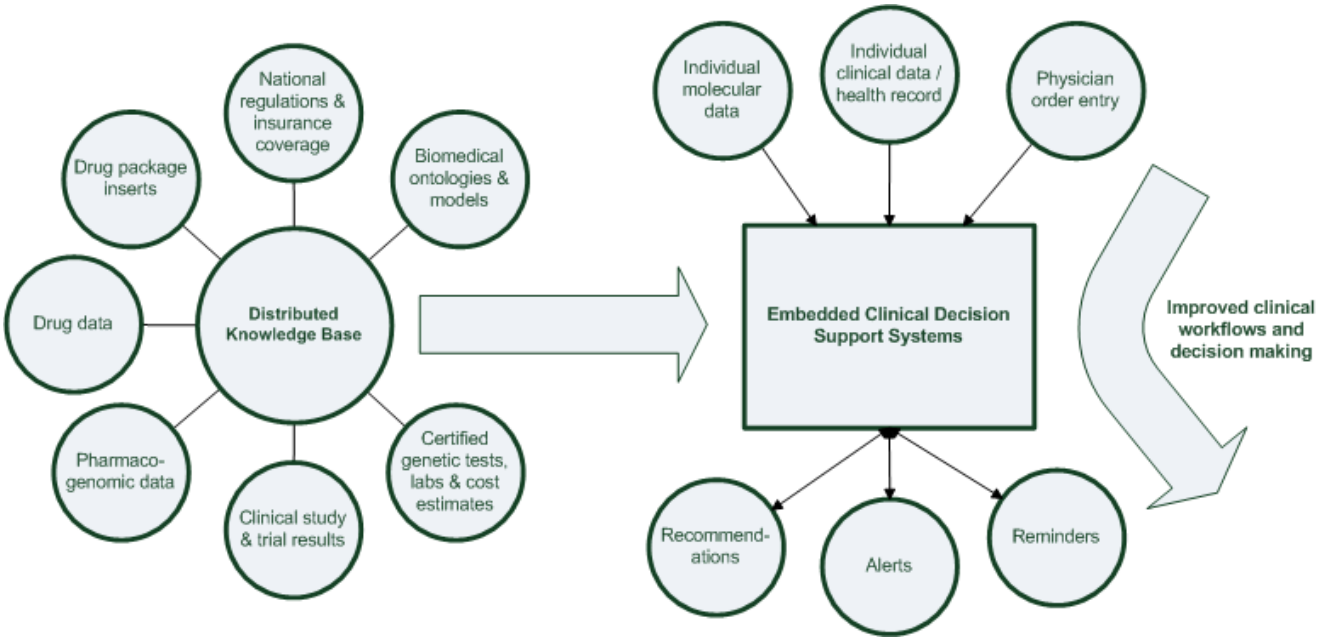

Figure 1: The components of an IT infrastructure for personalized pharmacotherapy. Relevant datasets such as genotype-phenotype associations or information about specific drugs are exposed publicly on the World Wide Web in Semantic Web formats. The datasets are interlinked, forming a distributed, yet coherent knowledge base. This knowledge base can then be used as the basis for the creation of clinical decision support systems which can reason over individual patient data when medical professionals are prescribing drugs.

We can conceptualize the ongoing discovery of a gene variant and its impact on medical outcomes as a continuous 'pharmacogenomic information pipeline' (table 4). Different stakeholders, datasets and technologies are associated with each level between initial experimental detection of a specific genetic variant (raw preclinical data) and the eventual clinical validation and deployment of genetic data in clinical decision making (established clinical rules that can be implemented in decision support systems).

Table 4: The pharmacogenomic information pipeline. We can conceptualize the translation of a pharmacogenomic finding from research (raw data) into practice (established clinical rules that can be implemented in decision support systems) as a continuum with several levels.

Level of establishment	*Current number of described loci with variation that have reached that level (order of magnitude)*
Level 1, Identification of variation: Experimental identification and validation of a human gene variant, submission to a gene variant database, annotation, description and uniform identification of the gene variant as a reference sequence or Locus Reference Genomic entry [54].	**4*10^7** (number of human RefSNP clusters in dbSNP)
Level 2, Clinical genotype-phenotype association: Clinical studies of phenotypes associated with genetic variants (e.g., drug response), deposition in genotype-to-phenotype databases.	**10^4** (estimate)
Level 3, Approval / recognition: Recognition of the clinical significance of the genotype-phenotype association by some authority, e.g., mention of impact of genetic variants in package inserts, approval of drugs for patients with specific genotype, clinical validation of companion diagnostics, modification of a national clinical guideline to contain genotype-based decision making.	**10^2** (estimate based on number of FDA product labels containing pharmacogenomic information [55].
Level 4, Significant clinical application: Application in clinical practice, possibly recognized as relevant by payers (reimbursement of diagnostic tests, requirement of testing for reimbursement of certain treatments). Implementation of pharmacogenomic guidelines in clinical decision support systems.	**10^1** (number of widely documented examples such as warfarin or herceptin)
Level 5, Surveillance: Monitoring of benefit, risk and cost associated with the implementation of a specific pharmacogenomic guideline in clinical practice	**?** (surveillance not yet established)

This pharmacogenomic information pipeline exhibits a steep decline in the number of genetic variants at the different stages from early clinical research to routine clinical application. It is currently not clear how long the transitions from one level to the next will take on average, and it is possible that the approval / recognition of pharmacogenomic guidelines by official authorities might be a major roadblock towards clinical adoption, especially if large-scale randomized clinical trials are seen as necessary prerequisites for approval. However, Altman *et. al* recently argued that, given the rapidly decreasing cost of genetic testing and the lack of potential harm relative to established practices, proof of *noninferiority* might be sufficient for initial implementation of pharmacogenomic guidelines in clinical practice [2]. In such a scenario, the number of pharmacogenomic findings potentially usable in clinical applications could increase rapidly over the coming decade, providing a challenge to clinicians who want to work with the best available evidence. An information infrastructure that makes findings of a specific quality directly available for use in clinical decision support systems, in that case, be critical to the adoption of pharmacogenomics in clinical practice. For example, the OWL ontology language and its automated reasoning functionality can be leveraged to describe subgroups of patients

with distinct pharmacogenomic characteristics, and to place individual patients into their appropriate pharmacogenomic groups (exemplified in figure 2).

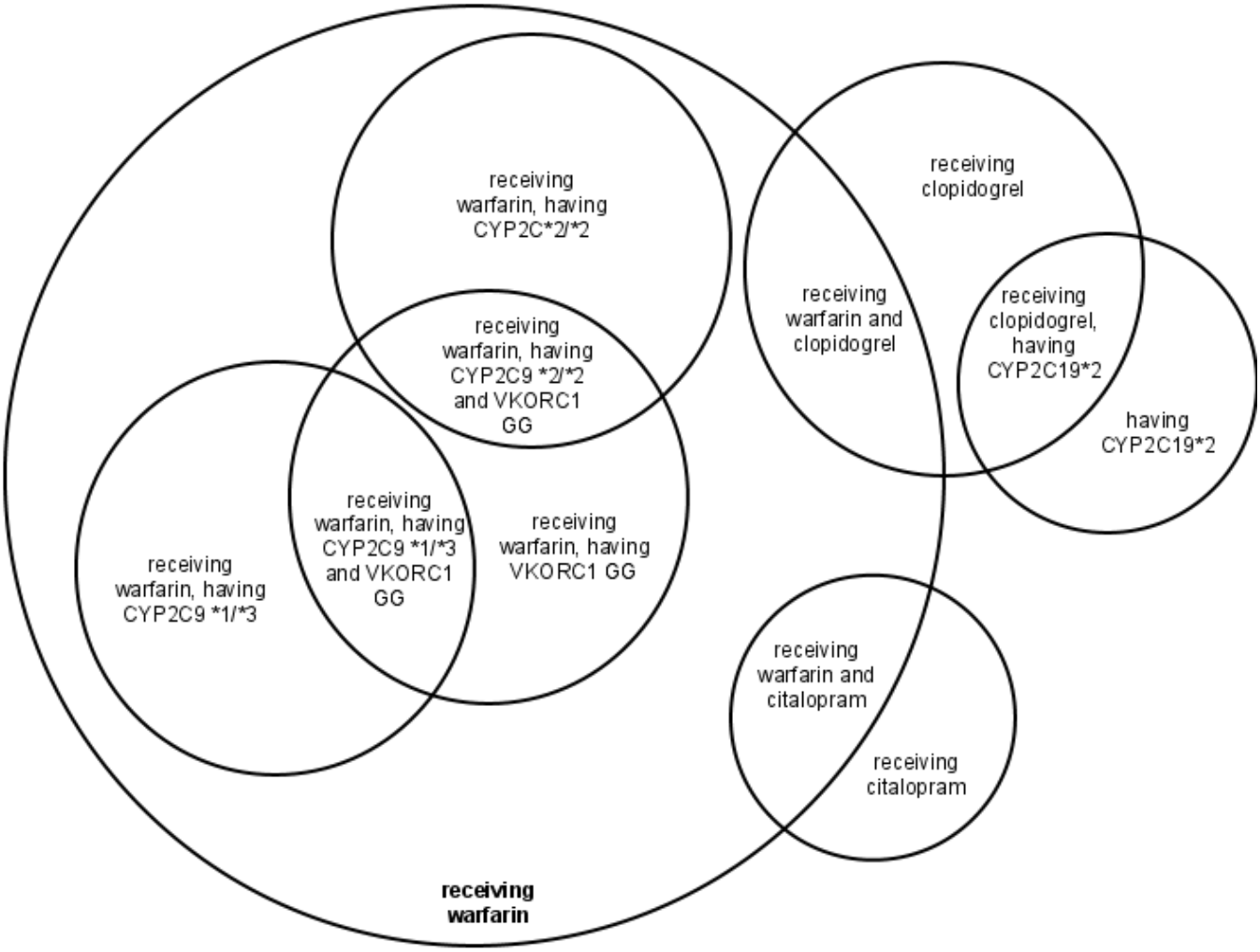

Figure 2: OWL can be used to describe groups and subgroups of patients with specific pharmacogenomic needs. Based on molecular data and basic clinical data, an OWL reasoner can automatically place each individual patient into their appropriate pharmacogenomic group. A decision support application can then provide medical professionals with alerts, reminders and recommendations for each specific patient group.

On the horizon

Several technologies and resources vital to the success of the approach described in this paper are currently under development. While the addition of new pharmacogenomic datasets to the Semantic Web is underway, substantial challenges remain in making the process of conversion sufficiently simple that non-experts can actively participate.

Further work is still required for strengthening the integration of Semantic Web technologies into established IT systems at hospitals, medical practices and pharmaceutical companies [56,57].

The work of the HL7 groups for clinical genomics [58] and clinical decision support [59] provide excellent starting points for such an endeavor.

Significant amounts of pharmaceutical data have been made available in RDF/OWL in recent years. Several pharmaceutical companies have internal Semantic Web projects for the purposes of data sharing within the enterprise. Furthermore, several large-scale projects based on Semantic Web formats for pre- and post-competitive information sharing in the pharmaceutical industry were launched recently. These include the SESL project of the Pistoia Alliance [60], the *Open Pharmacological Concepts Triple Store* (OpenPhacts) [61] and *Electronic Health Records for Clinical Research* (EHR4CR) [62] projects funded by the European Innovative Medicines Initiative. In the U.S., the eMERGE network [63] is a consortium of bio-repositories linked to electronic medical records data for conducting genomic studies organized by the National Human Genome Research Institute (NHGRI). Such initiatives could provide the critical mass necessary to build key parts of a pharmacogenomics information infrastructure.

The field of pharmacogenomics has grown significantly since the publication of the human genome, and advances in sequencing technology have resulted in an explosion in the amount of genomic data that is now available. As existing data sets are expanded and new sources of information are developed, the challenge of accessing and integrating information will continue to grow.

Semantic Web technologies have already been shown to help address challenges associated with pharmacogenomic data. The next decade will witness a convergence of advances in technology in both the laboratory and in computational infrastructure, which will present exciting opportunities to the field of pharmacogenomics and bring us closer to realizing the vision of personalized medicine.

Executive summary

- Pharmacogenomics has the potential to improve the effectiveness of health care and the development of new therapies.
- As pharmacogenomics becomes more complex, the therapies that it enables will depend on advanced decision support systems. These systems will utilize a semantic infrastructure for the biomedical sciences that is now being built.
- The datasets required for pharmacogenomics research and application in clinical practice are huge, distributed, heterogeneous and growing. The infrastructure that handles this must lower the cost of accessing and integrating such data.
- In order to discover associations between genes, gene variants, drugs, drug response, phenotypes, and diseases, the infrastructure should enable the seamless integration of data among relevant datasets.
- To realize the full potential of pharmacogenomics, the fruits of this technology need to be brought all the way from the research environment to the point of routine clinical decision making.

- Semantic Web technologies such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) are key standards for the creation of such an interoperable information infrastructure for translational, personalized medicine.
- Further key datasets for pharmacogenomics will be made available in RDF/OWL formats over the next three years. While challenges remain in integrating these technologies with existing IT systems at hospitals, medical practices and pharmaceutical companies, it is likely that first implementations in these settings will be deployed within the next five years.

References

1. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*. 19(12), 491-496 (2001).

2. Altman RB. Pharmacogenomics: “Noninferiority” Is Sufficient for Initial Implementation. *Clin Pharmacol Ther*. 89(3), 348-350 (2011).

3. Hafner JW Jr, Belknap SM, Squillante MD, Bucheit KA. Adverse drug events in emergency department patients. *Ann Emerg Med*. 39(3), 258-267 (2002).

4. King CR, Deych E, Milligan P, *et al.* Gamma-glutamyl carboxylase and its influence on warfarin dose. *Thromb. Haemost.* 104(4), 750-754 (2010).

5. Finkelman BS, Gage BF, Johnson JA, Brensinger CM, Kimmel SE. Genetic Warfarin Dosing. *J Am Coll Cardiol*. 57(5), 612-618 (2011).

6. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *N Engl J Med*. 360(8), 753-764 (2009).

7. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform*. 41(5), 706-716 (2008).

8. RDF Primer. (2004). Available from: http://www.w3.org/TR/rdf-primer/.

9. OWL Web Ontology Language Overview [Internet]. Available from: http://www.w3.org/TR/owl-features/.

10. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes M-D. Ontology-based knowledge discovery in pharmacogenomics. *Adv. Exp. Med. Biol.* 696, 357-366 (2011).

11. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*. 10(2), 153 -163 (2009).

12. Samwald M, Jentzsch A, Bouton C, *et al.* Linked open drug data for pharmaceutical research and development. *Journal of Cheminformatics*. 3, 19 (2011).

13. Klein TE, Chang JT, Cho MK, *et al.* Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *Pharmacogenomics J.* 1(3), 167-170 (2001).

14. The Pharmacogenomics Knowledge Base [PharmGKB] [Internet]. Available from: http://www.pharmgkb.org/.

15. GWAS Central - Home Page [Internet]. Available from: http://www.gwascentral.org/.

16. SNPedia [Internet]. Available from: http://www.snpedia.com/index.php/SNPedia.

17. OMIM Home [Internet]. Available from: http://www.ncbi.nlm.nih.gov/omim.

18. Home - dbGaP - NCBI [Internet]. Available from: http://www.ncbi.nlm.nih.gov/gap.

19. Webb AJ, Thorisson GA, Brookes AJ. An informatics project and online “Knowledge Centre” supporting modern genotype-to-phenotype research. *Hum. Mutat.* 32(5), 543-550 (2011).

20. GEN2PHEN Knowledge Centre [Internet]. Available from: http://www.gen2phen.org/.

21. GET-Evidence: About [Internet]. Available from: http://evidence.personalgenomes.org/about.

22. HuGENavigator [Internet]. Available from: http://hugenavigator.net/HuGENavigator/home.do.

23. Genetic Association Database [Internet]. Available from: http://geneticassociationdb.nih.gov/.

24. Genotator [Internet]. Available from: http://genotator.hms.harvard.edu/geno/.

25. GeneTests [Internet]. Available from: http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests.

26. Genetic Testing Registry [Internet]. Available from: http://www.ncbi.nlm.nih.gov/gtr/.

27. Luciano JS, Andersson B, Batchelor C, *et al.* The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics*. 2, S1 (2011).

28. Coulet A, Smaïl-Tabbone M, Napoli A, Devignes M-D. Ontology-based knowledge discovery in pharmacogenomics. *Adv. Exp. Med. Biol.* 696, 357-366 (2011).

29. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*. 10(2), 153 -163 (2009).

30. Laurila J, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker C. Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC Genomics*. 11(Suppl 4), S24 (2010).

31. Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. *J Biomed Inform*. 44(1), 87-93 (2011).

32. The Sequence Ontology - Index [Internet]. Available from: http://www.sequenceontology.org/.

33. de Matos P, Alcántara R, Dekker A, *et al.* Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.* 38(Database issue), D249-254 (2010).

34. Chemical Entities of Biological Interest (ChEBI) [Internet]. Available from: http://www.ebi.ac.uk/chebi/.

35. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc*. 18(4), 441-448 (2011).

36. RxNorm [Internet]. (2004). Available from: http://www.nlm.nih.gov/research/umls/rxnorm/.

37. McDonald CJ, Huff SM, Suico JG, *et al.* LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin Chem*. 49(4), 624-633 (2003).

38. Logical Observation Identifiers Names and Codes (LOINC®) — LOINC [Internet]. Available from: http://loinc.org/.

39. Osborne JD, Flatow J, Holko M, *et al.* Annotating the human genome with Disease Ontology. *BMC Genomics*. 10(Suppl 1), S6-S6.

40. Main Page - DO-Wiki [Internet]. Available from: http://do-wiki.nubic.northwestern.edu/index.php/Main_Page.

41. PATO:Main Page - OBOFoundry [Internet]. Available from: http://obofoundry.org/wiki/index.php/PATO:Main_Page.

42. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83(5), 610-615 (2008).

43. Rosse C, Mejino JLV Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform*. 36(6), 478-500 (2003).

44. MedDRA MSSO Welcome [Internet]. Available from: https://meddramsso.com/.

45. Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*. 11(10), 1467-1489 (2010).

46. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*. , 67-79 (2008).

47. Jonquet C, Lependu P, Falconer S, *et al.* NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources. *Web Semantics (Online)*. 9(3), 316-324 (2011).

48. Coulet A, Garten Y, Dumontier M, Altman RB, Musen MA, Shah NH. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *Journal of Biomedical Semantics*. 2, S10 (2011).

49. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*. 10 Suppl 2, S6 (2009).

50. A side effect resource to capture phenotypic e... [Mol Syst Biol. 2010] - PubMed - NCBI [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20087340.

51. Tari L, Anwar S, Liang S, Hakenberg J, Baral C. Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. *Pac Symp Biocomput*. , 465-476 (2010).

52. Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. *J Biomed Inform*. 43(6), 1009-1019 (2010).

53. Laurila J, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker C. Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC Genomics*. 11(Suppl 4), S24 (2010).

54. Dalgleish R, Flicek P, Cunningham F, *et al.* Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med*. 2(4), 24-24.

55. Genomics > Table of Pharmacogenomic Biomarkers in Drug Labels [Internet]. Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm.

56. Samwald M, Stenzhorn H, Dumontier M, Marshall MS, Luciano J, Adlassnig K-P. Towards an interoperable information infrastructure providing decision support for genomic medicine. *Stud Health Technol Inform*. 169, 165-169 (2011).

57. Heymans S, McKennirey M, Phillips J. Semantic validation of the use of SNOMED CT in HL7 clinical documents. *Journal of Biomedical Semantics*. 2, 2 (2011).

58. Farkash A, Neuvirth H, Goldschmidt Y, *et al.* A standard based approach for biomedical knowledge representation. *Stud Health Technol Inform*. 169, 689-693 (2011).

59. Clinical Decision Support [Internet]. Available from: http://www.hl7.org/Special/committees/dss/index.cfm.

60. SESL [Internet]. Available from: http://www.pistoiaalliance.org/workinggroups/sesl.html.

61. Open PHACTS [Internet]. Available from: http://www.openphacts.org/.

62. EHR4CR: Electronic Health Records for Clinical Research [Internet]. Available from: http://www.ehr4cr.eu/.

63. The eMERGE Network [Internet]. Available from: https://www.gwas.net/.

Reference annotations

** Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*. 19(12), 491-496 (2001).

** The Pharmacogenomics Knowledge Base [PharmGKB] [Internet]. Available from: http://www.pharmgkb.org/.

* Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*. 10(2), 153 -163 (2009).

* Luciano JS, Andersson B, Batchelor C, *et al.* The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics*. 2, S1 (2011).

Financial disclosure / Acknowledgements

The work of MS was funded in part by the Medical University of Vienna, as well as the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 257528 (KHRESMOI). The work of RRF was supported in part by U19 GM61388 (the Pharmacogenomics Research Network).