Published in final edited form as:

Pharmacogenomics. 2012 January ; 13(2): 201–212. doi:10.2217/pgs.11.179.

Semantically enabling pharmacogenomic data for the realization of personalized medicine

Matthias Samwald,

Section for Medical Expert & Knowledge-Based Systems, Center for Medical Statistics, Informatics, & Intelligent Systems, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria; Institute of Software Technology & Interactive Systems, University of Technology Vienna, Favoritenstrasse 9–11/188, A-1040 Vienna, Austria

Adrien Coulet,

LORIA – INRIA Nancy–Grand-Est, Campus Scientifique-BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France

Iker Huerga,

Elsevier, 1600 John F Kennedy Blvd. Suite 1800, Philadelphia, PA 19103-2899, USA

Robert L Powers,

Predictive Medicine, Inc., 37 Mansion Drive, Topsfield, MA 01983, USA

Joanne S Luciano,

Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Winslow 2143, Troy, NY 12180, USA

Robert R Freimuth,

Division of Biomedical Statistics & Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55901, USA

Frederick Whipple,

Genomics Education Initiative, 609 Sycamore Avenue, Fullerton, CA 92831, USA

Elgar Pichler,

Department of Chemistry & Chemical Biology, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA

Eric Prud'hommeaux,

World Wide Web Consortium/MIT, 32 Vassar Street, Cambridge, MA 02140, USA

Michel Dumontier, and

Department of Biology, School of Computer Science, Institute of Biochemistry, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1T 2T4, Canada

M Scott Marshall*

Department of Medical Statistics & Bioinformatics, Leiden University Medical Center/Informatics Institute, University of Amsterdam, Einthovenweg 20, 2333 ZC Leiden, The Netherlands

Abstract

© 2012 Future Medicine Ltd

*Author for correspondence: Department of Medical Statistics & Bioinformatics, Leiden University Medical Center/Informatics Institute, University of Amsterdam, Einthovenweg 20, 2333 ZC Leiden, The Netherlands, Tel.: +31 71 5269111
marshall@science.uva.nl

Understanding how each individual's genetics and physiology influences 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 critical 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

clinical decision support systems; knowledge representation; ontologies; personalized medicine; pharmacogenomics; translational medicine

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 could respond differently to the same therapy, owing 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 US FDA for which dosage recommendations based on specific combinations of genetic markers are provided (Table 1). In the USA alone, over 20 million prescriptions of warfarin are written per year to treat chronic anticoagulation indications including atrial fibrillation, deep vein thrombosis, pulmonary embolism and artificial heart valves. Over- or under-dosing can have serious consequences: warfarin is one of the leading causes of emergency care and drug-related hospitalization due to adverse drug events [1]. Genetic variations in two key genes have been found to strongly affect the toxicity of warfarin and consequently its initial recommended dose. First, the CYP2C9 enzyme 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 VKORC1 enzyme.

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 [2-4], 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 treatments that have been optimized 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 interactions that influence drug or target activity.

Semantic infrastructure for the biomedical sciences

In this article, we report on progress toward the development of a global infrastructure of semantic datasets and tools 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. In this article, 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. Internationalized Resource Identifiers (IRI) can be used to identify any entity, whether it is a hospital, a dosage regimen, a kind of drug, a kind of genetic variation or even a clinical report. An example of an IRI is that for warfarin as defined by the Drugbank database [5], but which is provided by the Bio2RDF project [6]. Entities identified by IRIs can be described in terms of their attributes and the relations they hold with other entities. The Resource Description Framework (RDF, [101]) 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 a particular drug has the brand name ‘Coumadin’ would be written as:

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

A statement that Coumadin (as defined by DrugBank entry) has warfarin as an ingredient (DB00682 as defined by ChEBI entry 10033) would be written:

```
<http://bio2rdf.org/drugbank_drugs:DB00682>  
<http://www.obofoundry.org/ro/ro.owl#has_part>  
<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 World Wide Web one of the most successful technologies in recent history.

Slightly more sophisticated than RDF is the Web Ontology Language (OWL) [102]. OWL is based on formal logic and can be used to capture general rules about the world (such as ‘every person has two biological parents’, ‘every *CYP2C9**2 allele has a thymidine nucleoside at position 430’) so that it becomes possible to answer questions that require automated reasoning [7,8]. OWL has already been used on many occasions to formally represent pharmacogenomics knowledge.

Linking open data

Linked Open Data (LOD) is a set of design principles that make it practical to share machine-readable information on the web. The set of services following these principles has come to be called the LOD cloud. This cloud is growing quickly, 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 healthcare and pharmacology problems. Heath *et al.* provide an extensive overview of the LOD best practices, datasets, tools and pitfalls [9].

At the core of the LOD cloud is RDF's use of IRIs (think URLs) for distributed extensibility. Using IRIs to represent, for example, 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 datasets.

LOD is designed to encourage the expression of linkages between data. The network basis for RDF allows us to express the interconnectedness of databases such as DailyMed [103], DrugBank [5] and the side effect resource (SIDER) [10], and to ask questions that reflect the real complexities of our domain. A wide variety of pharmaceutical datasets have been made openly available in Semantic Web formats by the LOD Data task force of the W3C [11].

Many tools for publishing and consuming linked data are openly available. For example, interlinking large datasets such as those in the LOD cloud was made possible by tools for automatic and semiautomatic link creation, such as Silk [12]. The D2R server [104] makes it possible to create wrappers around existing relational databases and to expose their contents as Linked Data. Virtuoso Open-Source [105] is a highly scalable web server, database and RDF store used in several biomedical applications. Sig.ma [106] is a search and visualization engine for Linked Data available on the public web. TripleMap [107] is an RDF/OWL-based web application for pharmaceutical, medical and clinical knowledge discovery.

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).

To create a comprehensive knowledge base regarding drugs and their pharmacogenomic properties, these data need to be combined with data regarding 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 5 years, the development of ontologies and related Semantic Web technologies have emerged as key elements with which to address this challenge. Together, they make it possible to both distribute and discover distributed data, as well as to retrieve and integrate it using applications built on Web standards. The capabilities of this approach have been demonstrated by developments such as:

- The Neurocommons knowledge base, a knowledge base comprising of a wide variety of biomedical datasets [13];

- The Translational Medicine Knowledge Base, a knowledge base with a focus on translational medicine [14];
- The Scientists' Networking System (SciNetS) of the RIKEN institute [108];
- BioPAX, an RDF/OWL based standard for sharing of molecular interactions and pathways, adopted by major interactomics databases [15];
- The Kidney and Urinary Pathway Knowledge Base [16].

Ontologies and terminologies play a critical role in data integration. They enable the use of well-defined, unambiguous terms to semantically annotate data, thereby providing the means by which one can query across different datasets that 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, as well as their relations, in a domain of discourse (such as pharmacogenomics). When an ontology is expressed in OWL, automatic reasoning can be performed in a predictable fashion. By ameliorating the complexity and heterogeneity of data representation, ontologies enable a separation of layers between pharmacogenomic knowledge, on the one hand, and both business rules of regulatory guidelines and clinician-facing application, on the other. The ontologically enabled knowledge layer then can be managed to track scientific advances independently of the other layers. Table 3 lists relevant ontologies and terminologies of interest to pharmacogenomics.

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. Conversely, Systematized Nomenclature of Medicine – Clinical Terminology (SNOMEDCT), one of the most widely employed general clinical terminologies, contains very few specific terms from the domains of genetics and pharmacogenomics. RDF/OWL representations of LOINC and SNOMED make it possible to merge data from these different coding schemes and ontologies into a comprehensive model. This model can include, for example, the LOINC test result 'chromosome analysis, peripheral blood' and the SNOMED CT finding 'biopsy result abnormal'. RDF/OWL thus 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.

OWL provides a flexible, standardized language to represent data and knowledge in many different ways. This flexibility makes it possible to represent the same types of biomedical findings in drastically different ways that can become difficult to align. Therefore, the potential to successfully integrate a newly created OWL ontology with the existing ontology infrastructure can be increased by following some basic design principles, such as adherence to an established top level ontology (e.g., the Basic Formal Ontology [BFO], [109]) and orientation on the original principles laid out by the Open Biomedical Ontologies (OBO) Foundry [110].

The BioPortal website and services offered by the US National Center for Biomedical Ontology can be used to find ontologies and to judge how popular each ontology is [17]. BioPortal includes a recommender service, to which a text or collections of concepts can be submitted and a list of suggested matching ontologies is generated in response.

To make newly created ontologies and datasets better discoverable, they should be documented on [thedatahub.org](https://www.thedatahub.org) [111], a popular website for sharing information on open datasets and their interrelations.

A great number of software tools is available for developing and using biomedical ontologies. Protégé 4 is a popular, openly available ontology editor [112]. TopBraid Composer [113] is a commercial alternative to Protégé, which also provides extensive functionalities for workflow creation, the creation of custom rules, data extraction, visualization and the creation of ontology-based end-user application. OWLIM is a highly scalable ontology repository and reasoning engine [114].

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 [18]. After extraction, entities and relations can be normalized with standard dictionaries and ontologies [19,20], and encoded in a structured format. Such normalized relations can subsequently be compared with other literature derived relations and to the content of other databases [21]. 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. The NLP community has begun to consider using full-text journal articles and alternative text resources [22,23]. This 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 [24], the automatic construction of databases such as SIDER [10], the completion of pharmacokinetics pathways [25], the creation of a Pharmacogenomic Relationships Ontology (PHARE) [26], and the extraction of mutation impacts on protein properties that were used to populate the Mutation Impact Ontology [27]. One limitation of NLP approaches is the lack of available text corpora where genes, drugs, phenotypes and their relationships are manually annotated. Such text corpora would help NLP researchers to train and evaluate their extraction models with improved efficiency. A second limitation is that most approaches extract relationships frequently observed in text, whereas some valid pharmacogenomics relationships are only rarely mentioned. Special methodologies need to be developed to correctly identify such relationships.

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 (Figure 1). Clinical decision support systems in the past often suffered from poor adoption and low impact on clinical practice for a variety of reasons, among them are:

- A lack of publicly available formally represented biomedical and clinical knowledge associating clinical findings with treatment recommendations;
- A lack of actionable patient data in electronic health record systems (or lack of such systems altogether);
- A lack of standardized ways of representing clinical data and findings, for example allergies, drug intolerance, phenotypic features, disease occurrence in close family members or radiological findings.

The envisioned infrastructure for pharmacogenomics could potentially circumvent some of these problems, because:

- Public databases associating genotypes with phenotypes (including clinically relevant phenotypes) are growing at an enormous pace.
- A single test (genotyping, sequencing) conducted at a single healthcare institution can yield a very large amount of clinically relevant data for each person. Potentially, the personal genetic make-up only needs to be tested once in the lifetime of each person and can then be stored and accessed through a secure electronic health record, and could be used to optimize future treatment decisions in a wide variety of unanticipated ways. This is in contrast to current approaches to electronic health records, where clinical findings accumulate slowly with clinical examinations over the years and need to be generated by different and geographically dispersed institutions (e.g., different laboratories, different clinical units).

It might prove easier to converge towards a standardized and interoperable way of representing personal genetic data (e.g., SNPs) than creating such representations for other types of clinical data and findings such as those mentioned above. While personal genetic data is large in size, it is relatively easy to formalize (in the most simplistic case as a string of nucleotide symbols).

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).

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 significant roadblock to 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 [28]. 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. We assert that in order to achieve the wide-scale adoption of pharmacogenomics in clinical decision support systems, an information infrastructure must be created, and must have the capacity to provide quality-assured research findings for use in such systems. 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).

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 nonexperts can actively participate.

Several parts of the infrastructure we describe have recently emerged or 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 nonexperts can actively participate. Many of the datasets listed in Table 2 have already been converted to RDF/OWL. Members of the W3C Health Care and Life Science Interest Group are currently working on converting and interlinking additional datasets, extending the Translational Medicine Ontology [14] to represent pharmacogenomic findings, and creating services for accessing the nascent infrastructure.

We expect that the maturation of this infrastructure will still take several more years. Different levels of maturity are needed for proofs-of-concept, research applications, and applications in routine clinical care. For obvious reasons, technology and datasets applied in a clinical setting will require a higher level of maturity than what is required for applications that do not directly impact patient care. Data intended for clinical applications will have to be carefully selected, curated and validated.

Further work is still required for strengthening the integration of Semantic Web technologies into established IT systems at hospitals, medical practices and pharmaceutical companies [29,30]. The work of the HL7 groups for clinical genomics [31] and clinical decision support [115] provide excellent starting points for such an endeavor.

Of course, the semantic infrastructure outlined in this paper is only one of many factors important for the realization of more targeted therapies. Besides the technical aspects, there are also social, legal and political challenges that need to be addressed.

While semantic technologies can help with integrating data and making it accessible to medical practitioners through clinical decision support tools, the requirement for expert committees to review and validate pharmacogenomic findings and to eventually formulate clinical guidelines should not be overlooked. It needs to be explored how the proposed infrastructure can be used to assist the creation and publication of such clinical guidelines.

Another open question is how to best display pharmacogenomic findings to medical doctors. Similar to many other type of clinical knowledge, findings based on genetic markers are often of a probabilistic nature (e.g., statements about risks of adverse events or altered probabilities of certain outcomes). It is well documented that medical doctors are not very good at interpreting such probabilistic findings [32]. User interfaces of decision support tools for enabling genomic medicine need to be capable of communicating such probabilistic findings in the most efficient and unambiguous way possible.

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 postcompetitive information sharing in the

pharmaceutical industry were launched recently. These include the SESL project of the Pistoia Alliance [116], the Open Pharmacological Concepts Triple Store (OpenPhacts) [117] and Electronic Health Records for Clinical Research (EHR4CR) [118] projects funded by the European Innovative Medicines Initiative (IMI). In the USA, the eMERGE network [119] 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. Future research will need to be directed towards investigating how different approaches taken by these initiatives (e.g., ontological modeling techniques) compare in their practical usefulness. In this paper, our objective was not to compare and contrast different approaches, but to demonstrate how current approaches can be combined to achieve a more powerful tool.

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 datasets 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.

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 [33]. 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 [28]. 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.

Acknowledgments

Financial & competing interests disclosure

The work of M Samwald 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 number 257528 (KHRESMOI). The work of RR Freimuth was supported in part by the NIH/NIGMS (U19 GM61388; the Pharmacogenomics Research Network). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:

■ of interest

■ ■ of considerable interest

1. Hafner JW Jr, Belknap SM, Squillante MD, Bucheit KA. Adverse drug events in emergency department patients. *Ann. Emerg. Med.* 2002; 39(3):258–267. [PubMed: 11867978]
2. King CR, Deych E, Milligan P, et al. Gamma-glutamyl carboxylase and its influence on warfarin dose. *Thromb. Haemost.* 2010; 104(4):750–754. [PubMed: 20694283]
3. Finkelman BS, Gage BF, Johnson JA, Brensinger CM, Kimmel SE. Genetic warfarin dosing. *J. Am. Coll. Cardiol.* 2011; 57(5):612–618. [PubMed: 21272753]
4. Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* 2009; 360(8):753–764. [PubMed: 19228618]
5. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* 2006; 34(Suppl. 1):D668–D672. [PubMed: 16381955]
6. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 2008; 41(5):706–716. [PubMed: 18472304]
7. Coulet A, Smail-Tabbone M, Napoli A, Devignes M-D. Ontology-based knowledge discovery in pharmacogenomics. *Adv. Exp. Med. Biol.* 2011; 696:357–366. [PubMed: 21431576]
8. Dumontier M, Villanueva-Rosales N. Towards pharmacogenomics knowledge discovery with the Semantic Web. *Brief. Bioinform.* 2009; 10(2):153–163. [PubMed: 19240125] ■ Describes a prototype of a Semantic Web representation of pharmacogenomic data.
9. Heath, T.; Bizer, C. Synthesis lectures on the Semantic Web: theory and technology. In: Hendler, J.; Van Harmelen, F., editors. *Linked Data: Evolving the Web into a Global Data Space*. Vol. 1. Morgan and Claypool Publishers; San Rafael, CA, USA: 2011. p. 1-136.
10. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 2010; 6:343. [PubMed: 20087340]
11. Samwald M, Jentzsch A, Bouton C, et al. Linked open drug data for pharmaceutical research and development. *J. Cheminform.* 2011; 3:19. [PubMed: 21575203]
12. Volz, J.; Bizer, C.; Gaedke, M.; Kobilarov, G. Silk – a link discovery framework for the web of data; Presented at: Proceedings of the 2nd Workshop Linked Data on the Web; 2009; p. 1-6.
13. Ruttenberg A, Rees JA, Samwald M, Marshall MS. Life sciences on the Semantic Web: the neurocommons and beyond. *Brief. Bioinform.* 2009; 10(2):193–204. [PubMed: 19282504]
14. 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. *J. Biomed. Semantics.* 2011; 2(Suppl. 2):S1. [PubMed: 21624155] ■ An ontological representation of translation medicine created by an international community.
15. Demir E, Cary MP, Paley S, et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* 2010; 28(9):935–942. [PubMed: 20829833]
16. Jupp S, Klein J, Schanstra J, Stevens R. Developing a Kidney and Urinary Pathway Knowledge Base. *J. Biomed. Semantics.* 2011; 2(Suppl. 2):S7. [PubMed: 21624162]
17. Musen M, Shah N, Noy N. BioPortal: ontologies and data resources with the click of a mouse. *AMIA Annu. Symp. Proc.* 2008:1223–1224. [PubMed: 18999306]
18. Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics.* 2010; 11(10):1467–1489. [PubMed: 21047206]
19. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb. Med. Inform.* 2008:67–79. [PubMed: 18660879]

20. Jonquet C, Lependu P, Falconer S, et al. NCBO resource index: ontology-based search and mining of biomedical resources. *Web Semantics (Online)*. 2011; 9(3):316–324. [PubMed: 21918645]
21. Coulet A, Garten Y, Dumontier M, Altman RB, Musen MA, Shah NH. Integration and publication of heterogeneous text-mined relationships on the Semantic Web. *J. Biomed. Semantics*. 2011; 2:S10. [PubMed: 21624156]
22. Cohen KB, Johnson HL, Verspoor K, Roeder C, Hunter LE. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*. 2010; 11:492. [PubMed: 20920264]
23. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J. Am. Med. Inform. Assoc*. 2011; 18(4):387–391. [PubMed: 21672908]
24. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*. 2009; 10(Suppl. 2):S6. [PubMed: 19208194]
25. Tari L, Anwar S, Liang S, Hakenberg J, Baral C. Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. *Pac. Symp. Biocomput*. 2010:465–476. [PubMed: 19908398]
26. Coulet A, Shah NH, Garten Y, Musen M, Altman RB. Using text to build semantic networks for pharmacogenomics. *J. Biomed. Inform*. 2010; 43(6):1009–1019. [PubMed: 20723615]
27. Laurila J, Naderi N, Witte R, Riazanov A, Kouznetsov A, Baker C. Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC Genomics*. 2010; 11(Suppl. 4):S24. [PubMed: 21143808]
28. Altman RB. Pharmacogenomics: ‘noninferiority’ is sufficient for initial implementation. *Clin. Pharmacol. Ther*. 2011; 89(3):348–350. [PubMed: 21326263]
29. 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*. 2011; 169:165–169. [PubMed: 21893735]
30. Heymans S, McKennirey M, Phillips J. Semantic validation of the use of SNOMED CT in HL7 clinical documents. *J. Biomed. Semantics*. 2011; 2(1):2. [PubMed: 21762489]
31. Farkash A, Neuvirth H, Goldschmidt Y, et al. A standard based approach for biomedical knowledge representation. *Stud. Health Technol. Inform*. 2011; 169:689–693. [PubMed: 21893835]
32. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. *Psychol. Sci. Public Interest*. 2007; 8(2):53–96.
33. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol*. 2001; 19(12):491–496. [PubMed: 11711191] ■■■ A key paper providing introductions to the concepts of personalized medicine.
34. 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*. 2001; 1(3):167–170. [PubMed: 11908751]
35. Webb AJ, Thorisson GA, Brookes AJ. An informatics project and online ‘knowledge centre’ supporting modern genotype-to-phenotype research. *Hum. Mutat*. 2011; 32(5):543–550. [PubMed: 21438073]
36. Mungall CJ, Batchelor C, Eilbeck K. Evolution of the sequence ontology terms and relationships. *J. Biomed. Inform*. 2011; 44(1):87–93. [PubMed: 20226267]
37. de Matos P, Alcántara R, Dekker A, et al. Chemical entities of biological interest: an update. *Nucleic Acids Res*. 2010; 38(Database issue):D249–D254. [PubMed: 19854951]
38. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inform. Assoc*. 2011; 18(4):441–448. [PubMed: 21515544]
39. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin. Chem*. 2003; 49(4):624–633. [PubMed: 12651816]
40. Osborne JD, Flatow J, Holko M, et al. Annotating the human genome with disease ontology. *BMC Genomics*. 2009; 10(Suppl. 1):S6–S6. [PubMed: 19594883]

41. 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.* 2008; 83(5): 610–615. [PubMed: 18950739]
42. Rosse C, Mejino JLV Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inform.* 2003; 36(6):478–500. [PubMed: 14759820]

Websites

101. RDF Primer. W3C recommendation. www.w3.org/TR/rdf-primer
102. OWL: Web Ontology Language Overview. www.w3.org/TR/owl-features
103. Dailymed. <http://dailymed.nlm.nih.gov/dailymed/about.cfm>
104. D2R server – publishing relational databases on the Semantic. Web. www4.wiwi.fu-berlin.de/bizer/d2r-server
105. Virtuoso Open-Source Wiki: Virtuoso Open-Source Edition. www.openlinksw.com/wiki/main/Main
106. Semantic Information MASHup. <http://sig.ma>
107. TripleMap. <http://triplemap.com>
108. RIKEN adopts Semantic Web as the data-release standard of its database-construction infrastructure. www.riken.jp/engn/r-world/info/release/press/2009/090331_2
109. Basic Formal Ontology. www.ifomis.org/bfo
110. Open Biological and Biomedical Ontologies: archive of original principles. www.obofoundry.org/crit_2006.shtml
111. The Data Hub. A registry of open knowledge datasets and projects. <http://thedatahub.org>
112. The Protégé Ontology Editor and Knowledge Acquisition System. <http://protege.stanford.edu>
113. TopBraid Composer. www.topquadrant.com/products/TB_Composer.html
114. OWLIM Ontotext. www.ontotext.com/owlim
115. Clinical decision support. www.hl7.org/Special/committees/dss/index.cfm
116. SESL. www.pistoiaalliance.org/workinggroups/sesl.html
117. Open PHACTS. www.openphacts.org
118. EHR4CR: electronic health records for clinical research. www.ehr4cr.eu
119. The eMERGE Network. www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page
120. The Pharmacogenomics Knowledge Base. www.pharmgkb.org ■■ A key resource of curated pharmacogenomic data.
121. GWAS Central. A genotype–phenotype association database. www.gwascentral.org
122. SNPedia. www.snpedia.com/index.php/SNPedia
123. OMIM. www.ncbi.nlm.nih.gov/omim
124. Home–dbGaP–NCBI. www.ncbi.nlm.nih.gov/gap
125. GEN2PHEN Knowledge Centre. www.gen2phen.org
126. GET-evidence. <http://evidence.personalgenomes.org/about>
127. HuGENavigator. <http://hugenavigator.net/HuGENavigator/home.do>
128. Genetic Association Database. <http://geneticassociationdb.nih.gov>
129. Genotator. <http://genotator.hms.harvard.edu/geno>
130. GeneTests. www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests
131. Genetic Testing Registry. www.ncbi.nlm.nih.gov/gtr
132. The Sequence Ontology. www.sequenceontology.org
133. Chemical Entities of Biological Interest. www.ebi.ac.uk/chebi
134. RxNorm. www.nlm.nih.gov/research/umls/rxnorm
135. Logical Observation Identifiers Names and Codes (LOINC®). <http://loinc.org>
136. DO-Wiki. http://do-wiki.nubic.northwestern.edu/index.php/Main_Page
137. PATO: OBOFoundry. http://obofoundry.org/wiki/index.php/PATO:Main_Page

138. MedDRA MSSO. <https://meddramsso.com>
139. Genomics table of pharmacogenomic biomarkers in drug labels. www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

Executive summary

- Pharmacogenomics has the potential to improve the effectiveness of healthcare 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 make use of a semantic infrastructure for the biomedical sciences that is now being built.
- The datasets required for pharmacogenomics research and application in clinical practice are large, 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 3 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 5 years.

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.

Figure 2. Web Ontology Language can be used to describe groups and subgroups of patients with specific pharmacogenomic needs

Based on molecular data and basic clinical data, a Web Ontology Language reasoner can automatically classify patients into the appropriate pharmacogenomic groups. A decision support application can then provide medical professionals with alerts, reminders and recommendations for each specific patient group.

Table 1
Range of expected therapeutic warfarin doses based on *CYP2C9* and *VKORC1* genotypes as described in a US FDA drug label

Genotype		<i>CYP2C9</i>					
		<i>*1/*1</i> (mg/day)	<i>*1/*2</i> (mg/day)	<i>*1/*3</i> (mg/day)	<i>*2/*2</i> (mg/day)	<i>*2/*3</i> (mg/day)	<i>*3/*3</i> (mg/day)
<i>VKORC1</i>	GG	5–7 [†]	5–7 [†]	3–4	3–4	3–4	0.5–2 [†]
	AG	5–7 [†]	3–4	3–4	3–4	0.5–2 [†]	0.5–2 [†]
	AA	3–4	3–4	0.5–2 [†]	0.5–2 [†]	0.5–2 [†]	0.5–2 [†]

Reproduced from the updated warfarin (Coumadin, Bristol–Myers Squibb, Princeton, NJ, USA) product label.

[†] Dosage recommendations deviate from standard dosage recommendations because of pharmacogenetic findings.

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

Name	Description
PharmGKB	A large database of curated knowledge and raw data about associations between genes, genetic variants, drug response and disease [34,120]
GWAS Central (formerly called HGVbaseG2P)	A database of GWAS that also provides summaries of study results [121]
SNPedia	A wiki-based platform containing information on phenotypes associated with SNP variants, population prevalence of genetic variants and SNP microarrays [122]
OMIM	Information about diseases with Mendelian inheritance, including references to the implicated genes [123]
dbGaP	Results of studies that have investigated the interaction of genotype and phenotype [124]
GEN2PHEN Knowledge Center	Integrated genotype-to-phenotype data with facilities for data annotation and user feedback [35,125]
GET-Evidence	A large database of automatically annotated and then manually curated information about the impact of genetic variations [126]
HuGE Navigator	Information on genetic variants, gene–disease associations, gene–gene and gene–environment interactions and evaluation of genetic tests [127]
GAD	Diseases associated with genetic variants [128]
Genotator	Aggregated gene–disease relationship data containing an integrated view over other datasets [129]
NCBI GeneTests	This resource concerns genetic tests used in diagnostic and genetic counseling [130]
The Genetic Testing Registry	A database (under development) about genetic markers and tests that enable their clinical exploration [131]

dbGaP: Database of Genotypes and Phenotypes; GAD: Genetic Association Database; GET: Genomes, environments, traits; GWAS: Genome-wide association studies; OMIM: Online Mendelian Inheritance in Man; PharmGKB: Pharmacogenomics Knowledge Database.

Table 3
Ontologies and terminologies of relevance for pharmacogenomics

Types of represented information	Name	Description
All of translational and personalized medicine	TMO	An ontology covering key aspects of the entire spectrum of translational and personalized medicine, developed by participants of the W3C Health Care and Life Science Interest Group [14]
PGx	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 [7]
PGx	PO	An ontology built from PharmGKB that includes biomedical measures and outcomes [8]
Mutation impact	Mutation Impact Ontology	An ontology designed to represent mutation impacts on protein properties resulting from an information extraction process [27]
Genotype	SO	Contains terms often used for the annotation of sequences and features, including detailed description of different types of sequence variations [36,132]
Chemical	ChEBI	ChEBI is a dictionary of molecular entities focused on small chemical compounds [37,133]
Chemical	RxNorm	Normalized names for clinical drugs, references to other terminologies [38,134]
Chemical, clinical	LOINC	An established coding system for clinical laboratory results. Contains many identifiers for results of genetic tests [39,135]
Phenotype	Disease Ontology	An ontology of human diseases [40,136]
Phenotype	PATO	An general ontology of qualities that can be used to describe phenotypes [137]
Phenotype	Human Phenotype Ontology	An ontology for phenotypic abnormalities encountered in human disease [41]
Anatomy	FMA	An ontology for the canonical, anatomical structure of an organism [42]
Safety	MedDRA	A terminology for safety reporting (mandated in Europe and Japan for safety reporting, standard for adverse event reporting in the USA) [138]

ChEBI: Chemical Entities of Biological Interest; FMA: Foundational Model of Anatomy; LOINC: Logical Observation Identifiers Names and Codes; MedDRA: Medical Dictionary for Regulatory Activities; PATO: Phenotypic Quality Ontology; PGx: Pharmacogenomics; PharmGKB: Pharmacogenomics Knowledge Base; PO: Pharmacogenomics Ontology; SO: Sequence Ontology; SO-Pharm: Suggested Ontology for Pharmacogenomics; TMO: Translational Medicine Ontology; W3C: World Wide Web Consortium.

Table 4
The pharmacogenomic information pipeline

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 [27]	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, for example, mention of impact of genetic variants in package inserts, approval of drugs for patients with specific genotype, clinical validation of companion diagnostics and modification of a national clinical guideline to contain genotype-based decision-making	10^2 (estimate based on number of US FDA product labels containing pharmacogenomic information [139])
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)

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.