



Evolution of an Adaptive Information System for Precision Medicine

Ana León^(✉), Alberto García S., Mireia Costa, Andrea Vañó Ribelles, and Oscar Pastor

PROS Research Center, Universitat Politècnica de València,
Camí de Vera, s/n, 46022 València, Spain
{aleon,algarsi3,opastor}@pros.upv.es,
{micossan,anvari}@etsii.upv.es

Abstract. Precision Medicine has emerged as a computational approach to provide a personalized diagnosis, based on the individual variability in genes, environment, and lifestyle. Success in such aim requires extensible, adaptive, and ontologically well-grounded Information Systems to store, manage, and analyze the large amounts of data generated by the scientific community. Using an existing adaptive information system (Delfos platform) supported by a conceptual schema and an AI algorithm, the contribution of this work is to describe how the system has been improved to address specific challenges regarding the clinical significance of DNA variants. To do so, the following topics are addressed: i) provide an ontologically-consistent representation of the problem domain; ii) improve the management of clinical significance conflicts; iii) ease the addition of new data sources; and iv) provide a scalable environment more aligned with the data analysis requirements in a clinical context. The aim of the work has been achieved by using a Model-Driven Engineering approach.

Keywords: MDE · Conceptual modeling · Information Systems · Precision Medicine

1 Introduction

Precision Medicine (PM) has emerged as a computational approach to interpret omics (e.g., proteomics, genomics, and metabolomics), facilitating their application to healthcare provision [2]. One of the pillars of the PM approach is the genetic diagnosis, that is based on determining the practical importance of each DNA variant according to its role in the development of disease (known as clinical significance). There are different public databases that provide interpretations of the clinical significance of variants (i.e. variant interpretations) such as ClinVar (www.ncbi.nlm.nih.gov/clinvar/), Ensembl (www.ensembl.org/index.html), ClinGen (www.clinicalgenome.org) and CIViC (www.civicdb.org/home).

Even though the mentioned databases are an excellent source of information, the interpretation of the clinical significance that they provide is a challenging

process that may significantly affect diagnosis and clinical care recommendations [4]. Our experience with these repositories has allowed us to identify two main problems: i) lack of a clear representation of the clinical significance at phenotype level; and ii) a generic and sometimes not very precise identification of conflicts between interpretations. The consequence of these problems are explained in detail in Sect. 2.

Adaptive Information Systems (AIS) are key to overcome these challenges, by extending and adapting their functionality to the dynamism of the domain, presenting the available evidence with a well-grounded ontological basis, and providing automated algorithms to properly handle conflicts. The contribution of this work is to describe how using a Model Driven Engineering (MDE) approach, the above mentioned problems can be solved to improve an existing genomic platform, Delfos platform, to: i) consistently represent what a variant interpretation is; ii) allow the efficient management of conflicts between interpretations; and iii) provide a consistent environment for the precise evaluation of the clinical significance of DNA variants in the context of an efficient genomic data management.

To this aim, the work is organized as follows: Sect. 2, describes in detail the problem and the consequences from an information Systems Engineering perspective. Section 3 presents our proposed solution. Finally, Sect. 4 concludes and discusses future research directions.

2 Clinical Significance and Conflict Management

2.1 The Clinical Significance

The clinical significance is the practical importance of a variant effect (e.g., benign, pathogenic, or uncertain significance). The clinical significance of each variant is interpreted by experts, after the review and evaluation of the available evidence that supports the association of the variant with a phenotype (trait or disease). Different public and nonpublic databases provide interpretations of the clinical significance of variants, as introduced in Sect. 1.

A DNA variant can be interpreted multiple times by different experts and for different phenotypes. To help experts assess the clinical interest of a variant, an aggregate clinical significance is usually provided by these databases, which is useful to determine if the different interpretations are concordant or discordant. For example, the variant $c.986A > C$ has been interpreted in ClinVar by 13 experts for different phenotypes (e.g. glycogen storage disease, GBE1-Related disorders, and polyglucosan body disease) [1]. As all the experts consider the variant as pathogenic in all the interpretations, and for all the phenotypes, the aggregate clinical significance is pathogenic.

Nevertheless, the complexity of human disease implies that the effect of a variant may be different for different phenotypes. In such cases, the databases do not compute a precise aggregate, and the user must review and analyse each of the experts' interpretations to identify the correct role of the variant for each phenotype. This frequently conforms a tedious, manual, and prone-to-error

working process that diminishes the added value of Information Systems for the development of an efficient PM. Nevertheless, the higher impact of this approach occurs when conflicts between interpretations are analyzed.

2.2 Conflicting Interpretations

The conflicts between interpretations arise when experts disagree about the role of a variant in the development of disease. In general, interpretations have a high degree of concordance [7]. However, as knowledge about the mechanisms of disease evolves, the existence of conflicts in the interpretation of variants over time is not uncommon [6].

As mentioned, the different interpretations of a variant are typically aggregated into a “global” clinical significance. As a consequence, a variant that has been interpreted as disease causing for a given phenotype, and as not disease causing or uncertain for another, could be considered as having conflicting interpretations. These variants are more likely to be discarded from genetic diagnosis since they are considered as conflicting, although their exclusion could lead to missing important information.

Thus, the precise analysis and treatment of the conflicts is a key feature of any information system that integrates data from different sources.

3 Extending an AIS by Adding the Clinical Actionability

In the PROS Research Center (<http://www.pros.webs.upv.es/>), we have developed an AIS, called the Delfos platform, ontologically supported by the Conceptual Schema of the Human Genome (CSHG). [3,5] and a deterministic classification AI algorithm.

The aim of Delfos is to ease the management of the genetic data with clinical purposes. Thanks to the ontological support of the CSHG, Delfos can be extended to include new functionality, and consequently can be adapted to any change in the domain.

Initially, the CSHG modeled variants so that they can be associated with multiple clinical interpretations (see Fig. 1). Each variant (Variation class) may have multiple clinical interpretations provided by the scientific community (Significance class) for each Phenotype. Interpretations are described by the “ClinicalSignificance” and the “levelOfCertainty” attributes. The “ClinicalSignificance” determines the practical importance of the variant. The “levelOfCertainty” represents the relevance of the evidence used by each expert to assess that importance.

Nevertheless, in the context of an advanced genomic management platform, the aggregate clinical significance approach, followed by most of the genomic sources, is not useful because of the problems and uncertainties mentioned in Sect. 2. This led to the need of providing a better solution.

To this aim, we have followed a MDE approach with the following steps: i) an ontological characterization of the main concepts, ii) an extension of the CSHG

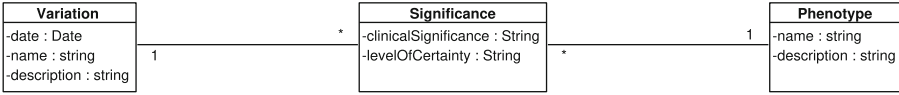


Fig. 1. Clinical Significance in the CSHG

to represent the new knowledge, and iii) an application of changes to make a new version of the Delfos platform. MDE promotes the systematic use of models in order to raise the level of abstraction at which software is specified, increasing the level of automation in software development, what we consider to be the most appropriate approach according to the context and aim of this work.

3.1 Ontological Characterization

As mentioned in Sect. 2.1, the clinical significance is the practical importance of a variant effect, commonly associated with a phenotype. The impact of this effect is characterized according to terms such as Pathogenic (variants that cause a disorder), Protective (variants that decrease the risk of a disorder) and Uncertain significance (variants with insufficient or conflicting evidence about their role in disease).

To help assess the degree of concordance between interpretations, databases compute an aggregate clinical significance, but without specifying which one corresponds to each phenotype. This means that the treatment of conflicts are reduced to a limited number of terms, excluding potentially relevant combinations.

3.2 Evolution of the Conceptual Schema

The different types of clinical significances can be grouped according to their likelihood of being the cause of a potentially damaging phenotype, or providing protection against one. Clinical significances related to drug or treatment responses are special cases since their definition does not specify if the effect is positive or negative.

Using this approach as basis, we propose to create an aggregate value for each phenotype associated to a variant, by grouping the different interpretations into a new conceptual entity that we have called “clinical actionability”. Therefore, instead of having a general term for each variant (an approach whose limitations have been stated in Sect. 2), the information system would provide a set of terms that allows a more precise assessment of a variant effect, according to the different phenotypes that have been studied. This approach is more aligned with the data analysis requirements in the context of a clinical practice. To represent this new knowledge, and provide the Delfos platform with the new functionality, the conceptual schema on which the information system is based must be modified.

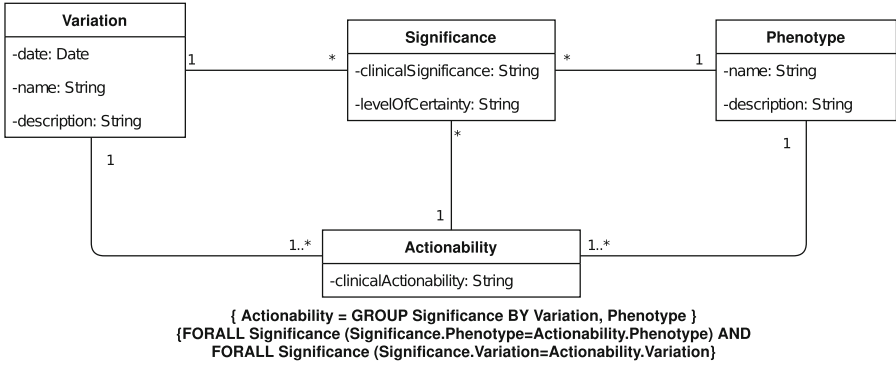


Fig. 2. Conceptual schema to represent the clinical actionability

Figure 2 represents the new Actionability class, associated with the Variation, Phenotype, and Significance classes. The clinicalActionability attribute (Actionability class) represents the practical importance of the variant effect. For each phenotype of a variant (Phenotype class), the clinical actionability is calculated as an aggregate of the different clinical significances (Significance class) provided by experts. Only one clinical actionability is allowed for each Variation-Phenotype pair (represented as a constraint to ensure data integrity).

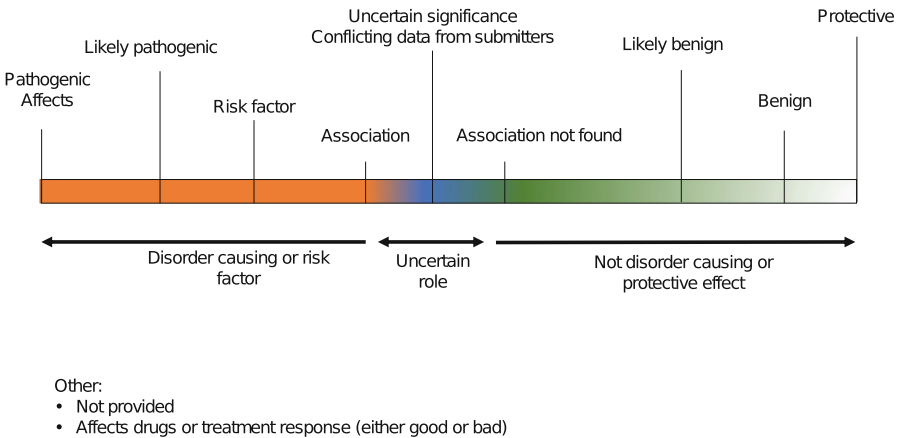


Fig. 3. Distribution of the different clinical significance types according to their likelihood of being the cause of a potentially damaging phenotype.

Once the conceptual schema is defined, the next step is to specify how the clinical actionability is calculated. Using the likelihood distribution shown in Fig. 3, we have defined the following terms to describe the different clinical actionability types:

- Disorder causing or risk factor: The variant is the cause of the phenotype, or increases the likelihood of presenting it. This group includes the interpretations whose clinical significance is pathogenic, likely pathogenic, affects, risk factor, or association.
- Not disorder causing or protective effect: The variant is not the cause of the phenotype, or provides a protective effect against it. This group includes the interpretations whose clinical significance is benign, likely benign, association not found, or protective.
- Affects drugs or treatment response: The variant affects the sensitivity or response to the specified drug or treatment. This group includes the interpretations whose clinical significance is drug response or confers sensitivity.
- Uncertain role: The role of the variant in the development of the phenotype is not clear. This group includes the interpretations whose clinical significance is uncertain significance, or when conflicts between interpretations are present.
- Not provided: The variant does not have interpretations and as a consequence the clinical significance is unknown.

Conflicts between interpretations occur when there is less than 75% of agreement in the role of the variant, regarding the development of the associated phenotype. This decision has been taken to avoid situations where an old or not reliable interpretation contradicts the major agreement of the scientific community. Conflicts occur in the following situations:

- Presence of interpretations whose clinical significance belongs to the disorder causing or risk factor group, and to the not disorder causing or protective effect group.
- Presence of interpretations whose clinical significance belongs to the disorder causing or risk factor group, and to the uncertain role group.
- Presence of interpretations whose clinical significance belongs to the not disorder causing or protective effect group, and to the uncertain role group.

Interpretations with no clinical significance provided are not considered for the identification of conflicts. For example, if there are three interpretations for the same phenotype - one of them pathogenic, another one benign, and the third one without the clinical significance specified - only the pathogenic and benign interpretations will be considered, resulting in the presence of conflicts. As a consequence, in this example, the clinical actionability of the variant will belong to the uncertain role group.

Despite the low impact of the changes at the conceptual-model level, the implications for the analytical capabilities of the information system are relevant. The impact of these changes are: i) Abstraction of the different variant effects according to their likelihood of being disease causing or protective, ii) Possibility of evaluating the clinical impact of variants for each associated phenotype, and iii) Decrease of the effort required to add new data sources that use different terms to classify the clinical significance. These changes in the conceptual schema, have been translated into changes in the implementation of the information system that supports the Delfos platform.

3.3 New Delfos Version

The AIS that constitutes the core of the Delfos platform has three main modules:

1. The extraction and transformation module connects to the databases that provide the input data to the system.
2. The identification module is based on a deterministic AI classification algorithm that evaluates the input data, and uses the relationships between the concepts of the CSHG to identify clinically relevant variants.
3. The visualization and exploitation module provides the Graphic User Interface required to query and visualize the knowledge stored in the database that serves as internal data storage.

The main changes affect the AI algorithm, and the way the new knowledge is visualized and accessed by the final user. The rules used to define the different clinical actionability groups, and the criteria required to identify conflicts between interpretations, have been added to the AI algorithm. These changes improve its capability of identifying relevant variants. The internal data storage has been modified to store this new knowledge, according to the specifications of the conceptual schema. Finally, the visualization and exploitation layer has been adapted to provide the required usability.

Thanks to the above mentioned changes, the Delfos platform has been improved to correctly address the problems mentioned in Sect. 2. Using the approach presented in this work, the Delfos platform is able to identify which phenotypes have real conflicts between interpretations, and considers that the effect of the variants could be relevant in other cases. If this information were missing in a genetic analysis, the diagnosis and treatment of patients would be severely affected.

4 Conclusion and Future Work

AIS are key to provide the technological support required to developing correct and accurate genetic diagnosis in the dynamic context of PM. In this work, we have identified two main challenges that led to the need of improving an existing information system (the Delfos platform). The first challenge was the lack of a clear characterization of the variant's clinical significance interpretation at phenotype level; and the second challenge was a generic and sometimes not very precise identification of conflicts between interpretations.

Since Delfos is an AIS supported by a conceptual model and an AI algorithm, we have improved the system by using a MDE approach to: i) consistently represent what a variant interpretation is; ii) allow the efficient management of conflicts between interpretations; iii) ease the integration of interpretations coming from different data sources; and iv) provide a consistent environment aligned with the data analysis requirements in the context of a clinical practice.

Genomics knowledge is under constant evolution. Therefore, the Delfos platform must be frequently updated to adapt to the dynamism of the domain. The

main advantage of using and AIS platform is that its extension can be done by reusing what has already been developed, focusing on evolving only the parts that have changed, and reducing the development effort required.

Acknowledgements. This work was supported by the Spanish State Research Agency [grant number TIN2016-80811-P]; and the Generalitat Valenciana [grant number PROMETEO/2018/176] co-financed with European Regional Development Fund (ERDF).

References

1. ClinVar Variant Details (vcv000002777.10). <https://www.ncbi.nlm.nih.gov/clinvar/variation/2777/>. Accessed 20 Oct 2020
2. Duffy, D.J.: Problems, challenges and promises: perspectives on precision medicine. *Briefings Bioinform.* **17**(3), 494–504 (2016). <https://doi.org/10.1093/bib/bbv060>
3. García, S.A., Palacio, A.L., Reyes Román, J.F., Casamayor, J.C., Pastor, O.: Towards the understanding of the human genome: a holistic conceptual modeling approach. *IEEE Access* **8**, 197111–197123 (2020). <https://doi.org/10.1109/ACCESS.2020.3034793>
4. Pepin, M.G., Murray, M.L., Bailey, S., Leistriz-Kessler, D., Schwarze, U., Byers, P.H.: The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories. *Genet. Med.* **18**(1), 20–24 (2016). <https://doi.org/10.1038/gim.2015.31>
5. Reyes Román, J.F., Pastor, Ó., Casamayor, J.C., Valverde, F.: Applying conceptual modeling to better understand the human genome. In: Comyn-Wattiau, I., Tanaka, K., Song, I.-Y., Yamamoto, S., Saeki, M. (eds.) *ER 2016. LNCS*, vol. 9974, pp. 404–412. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46397-1_31
6. Shah, N., et al.: Identification of misclassified ClinVar variants via disease population prevalence. *Am. J. Hum. Genet.* **102**(4), 609–619 (2018). <https://doi.org/10.1016/j.ajhg.2018.02.019>
7. Yang, S., Lincoln, S.E., Kobayashi, Y., Nykamp, K., Nussbaum, R.L., Topper, S.: Sources of discordance among germ-line variant classifications in ClinVar. *Genet. Med.* **19**(10), 1118–1126 (2017). <https://doi.org/10.1038/gim.2017.60>