A Conceptual Modeling-based Journey into Variant Interpretation: From Unpacking to Operationalization

Mireia Costa[1,*]

[1]PROS Group, Valencian Research Institute (VRAIN), Universitat Politècnica de València

Abstract

The process of determining the role that a DNA variant has on an individual's health status is known as variant interpretation. In the era of precision medicine, variant interpretation has become essential in clinical decision-making. Despite its relevance, variant interpretation is often criticized for being qualitative and open to expert interpretation. Clinicians argue that more concrete definitions are needed to systematize variant interpretation. This Ph.D. Thesis intends to cover these needs by providing artifacts to unpack and operationalize the variant interpretation process. The unpacking process has been achieved through conceptual modeling due to its proven effectiveness in similar domains. The operationalization comes from a method that defines variant interpretation in a set of small and well-defined steps. The design of these artifacts has been guided by the Design Science Methodology as proposed by Roel Wieringa. We envision that the unpacking and operationalization proposed in this research will result in a more explainable, reproducible, and reliable variant interpretation.

Keywords

Conceptual Modeling, Variant Interpretation, Precision Medicine

1. Introduction

In recent years, a new paradigm in medicine has revolutionize patient diagnosis and treatment. This paradigm, known as **precision medicine**, puts the spotlight on each patient's uniqueness and seeks to deliver the most accurate clinical actions based on each individual's characteristics [1], rather than using the traditional *one-size-fits-all* approach [2].

Despite the fact that individuals share about 99% of their DNA sequence, our DNA remains considered one of our most distinctive characteristics. The reason is that these minor differences in our DNA account for both natural human variability and susceptibility to certain diseases or an altered response to standard treatments. These differences in the DNA sequence are known as **DNA variants**. Because of their relevance to human health, one of the primary goals of precision medicine is to determine how DNA variants affect each individual's health status. This process is referred to as **variant interpretation**.

Variant interpretation is a **knowledge-driven** process that involves weighting many sorts of evidence scattered across thousands of data sources regarding the DNA variants that are being interpreted [3]. Examples include the variant's frequency among different populations and

CAiSE 2024 Doctoral Consortium

[*] Corresponding author.

[†] These authors contributed equally.

micossan@vrain.upv.es (M. Costa)

0000-0002-8614-0914 (M. Costa)

whether other experts have already linked the variant to a disorder [4]. Geneticists and clinical professionals are still arguing about how to weigh this evidence to accurately determine the impact of a variant on our health status. To address this issue, several authors have developed **variant interpretation guidelines**, a series of instructions designed to guide the interpretation process by determining whether or not a variant meets specific criteria.

Geneticists have widely embraced these recommendations into their daily practice [5], transforming the traditional ad-hoc variant interpretation process into a **rule-driven** one. Variant interpretation guidelines have been a significant step forward in the variant interpretation domain. However, they are far from being a definitive solution, as they have frequently been criticized for their qualitative nature and lack of specificity [6, 7]. Indeed, experts express the concern that despite the use of these guidelines, variant interpretation is still a **highly variable** process whose practical application is often left open to expert interpretation [8], thus hampering systematization.

Clinical experts state that establishing more specific definitions is essential for standardizing variant interpretation [9]. This Ph.D. Thesis takes as a challenge overcoming the problems in the domain by defining artifacts to: i) unpack the variant interpretation domain by providing the necessary definitions, and ii) operationalize the interpretation process by providing the means for facilitating its systematization. Conceptual Modeling has proven to be effective in achieving high levels of clarification and standardization in several clinical domains [10, 11, 12] and, thus, it will guide this Thesis' *journey* of tackling the problems in the variant interpretation domain. Therefore, this Ph.D. Thesis benefits from traditional Information System Engineering (ISE) approaches, but applies them to a non-traditional domain.

The remainder of the paper is organized as follows. Section 2 deepens on the variant interpretation challenges. Section 3 summarizes the objectives of this Thesis, and Section 4 describes the methodology followed to achieve such objectives. Finally, Section 5 describes the current state of the Thesis and Section 6 concludes the paper with a future outlook.

2. Problem statement

Variant interpretation is a knowledge-driven, rule-driven, and highly variable process. The factors contributing to variant interpretation being a challenging domain are encompassed in one of these characteristics.

As a **knowledge-driven** process, variant interpretation relies on both *explicit* and *tacit* knowledge [13]. Explicit knowledge is often encoded as a knowledge base that stores all the relevant information of the domain. Unfortunately, in this domain, there is no comprehensive source of information that is sufficient to provide all the needed evidence. Instead, the evidence is scattered across thousands of data sources, each with its unique content, structure, and terminology [3]. Some examples of widely-used data sources are ClinVar [14] and LOVD [15]. Against this backdrop, clinical experts face the problem of selecting which data sources to focus on, how to identify the relevant information in them, and how to exploit it to perform their assessments [16]. Tacit knowledge also plays a significant role. Even if they are unaware, experts will use their expertise and experience as an essential contributor to the interpretation process [17]. The combination of the data chaos affecting explicit knowledge and the ethereal nature of tacit knowledge often makes the evidence used for interpretation difficult to trace and potentially causes the same piece to be interpreted differently by different experts [18].

As a **rule-driven** process, variant interpretation decision making is conducted by following the instructions defined in variant interpretation guidelines. An example of an interpretation guideline are the ACMG-AMP 2015 guidelines [19]. Instructions are intended to provide clear and actionable steps to achieve the desired result. However, variant interpretation guidelines are criticized for providing vague definitions that allow a high degrees of subjectivity and uncertainty. In this scenario, inconsistent interpretations among experts become common. Consider, for instance, a real case where an initial assessment in prenatal care revealed that an unborn child is at high risk of developing Muscular Dystrophy disorder. However, a different team of experts later repeated this assessment and determined that it was incorrect [20]. Because these families often have to make decisions about pregnancy management within a short time span, the improperly classified variant could have had irreversible repercussions. Furthermore, the more complex the disorder (e.g., cancer), the more inconsistencies in variant interpretation usually occur [21]. The qualitative nature of variant interpretation guidelines can have serious consequences in healthcare applications. Consequently, clinical experts argue that these guidelines are too qualitative in nature and more concrete definitions are needed.

Finally, as a **high variable** process, it is highly dependent on the expert performing the interpretation. This is a direct consequence of the lack of specificity in both the evidence used and the rules applied. This variability causes variant interpretation to lack of traceability and reproducibility, which are dangerous characteristics for a process with such a significant impact on clinical decision making. In an effort to reduce variability, several tools have been developed to automate and the variant interpretation, hoping that this will enhance systematization. These tools include Varsome [22] and InterVar [23]. However, as with human application, the qualitative nature and insufficient specificity that characterize variant interpretation cause the various tools to make different assumptions and thus also interpret variants in discordant ways.

From all the above, it is clear that the variant interpretation domain requires disambiguation and systematization to avoid putting patients' health at risk and to achieve its successful application on a clinical daily basis [9]. Addressing these problems is the main objective of this Thesis, as we build on the following section.

3. Objectives

The problems hindering variant interpretation can be summarized as the lack of precise definitions and systematization. Therefore, the main objective of this Thesis is to provide the means for **unpack and operationalize the variant interpretation process**. This objective has been divided into three concrete goals:

– **G1**. Study the variant interpretation domain. This goal aims to identify all of the constructs involved in the variant interpretation process, the current challenges, and existing approaches tackling any identified issues. The following research questions (RQ) will guide the goal achievement:

RQ1. How is variant interpretation performed?

RQ2. What are the main challenges hampering variant interpretation?

RQ3. What existing approaches tackle any of the identified challenges?

– **G2**. Design artifacts to unpack and operationalize variant interpretation. Here, we aim to propose solutions to the challenges identified in the previous goal. By overcoming the current issues in the domain, we will provide the means for achieving a precise definition of the domain constructs (unpacking) and facilitate its systematic application (operationalization). The following RQ are to be answered:

RQ4. How to unpack the variant interpretation process?

RQ5. How to operationalize the variant interpretation process?

– **G3**. Validate the designed artifacts. In this goal, we will validate that the designed artifacts effectively solve the identified problems by answering the following RQ:

RQ6. To which extent the designed artifacts allow to unpack the variant interpretation process?

RQ7. To which extent the designed artifacts allow one to operationalize the variant interpretation process?

4. Research methodology

This Thesis follows the research method of *Design Science* as proposed by Roel Wieringa [24], which consists of designing and investigating artifacts in a context to provide solutions for a specific problem. In this research, the artifacts will be the proposed *solutions to unpack and operationalize* the variant interpretation process, with *precision medicine* as the context.

Following the recommendations of Wieringa for practical problems, we have followed a design cycle of three phases:

– **Problem investigation**: The problem is characterized and its context is described.

– **Treatment design**: The artifacts to address the problem under investigation are provided.

– **Treatment validation**: The adequacy of the proposed artifacts is validated in the selected context.

The design cycle typically has a fourth stage called **Design implementation** that focuses on the technological transfer of the designed artifacts into a real-word scenario. This stage is outside the scope of this Thesis.

Figure 1 summarizes how the objectives proposed in this Thesis align with the Design Science methodology.

5. Current status and contributions

This Thesis has been underway for 2.5 years. Until now, efforts have been focused on the problem investigation and treatment design stages. Below, we discuss the current status of these stages, focusing on how each RQ has been answered and the artifacts that have been generated.

– *RQ1. How is variant interpretation performed?*

The answer to this RQ began with a thorough understanding of genomics, which is the field of study that focuses on understanding the DNA. Following that, we investigated which are the steps typically used to interpret a variant. This investigation was supported by various clinical experts who are collaborating on research projects currently undergoing in our research group.

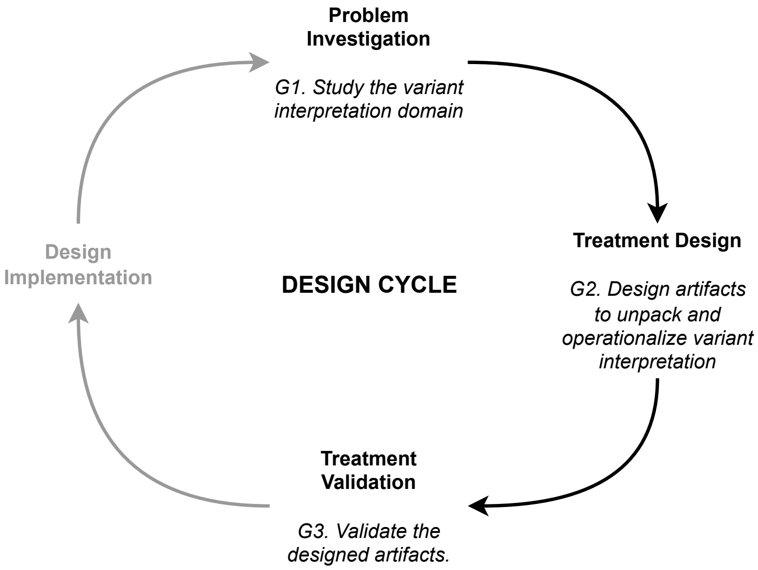
Problem Investigation
G1. Study the variant interpretation domain
Treatment Design
G2. Design artifacts to unpack and operationalize variant interpretation
Treatment Validation
G3. Validate the designed artifacts.
Design Implementation
DESIGN CYCLE

Figure 1: Design Science application in this Ph.D. Thesis.

As a result of this research, a BPMN model that represents the variant interpretation process *as-is* has been generated.

– *RQ2. What are the main challenges hampering variant interpretation?*

Section 2 already gives an outline of the issues that hinder variant interpretation. The problems were identified by analyzing 13 scientific articles that, carrying out different experiments, discussed the origins of variant interpretation conflicts among experts. From this analysis, the underlying problems were identified and brought up to light.

– *RQ3. What existing approaches tackle any of the identified challenges?*

There are some approaches that partially address the problems affecting variant interpretation. First, there are data sources that collect information about variant interpretations performed by clinical experts worldwide, aiding in the problem of explicit knowledge dispersion. However, they are neither complete nor concordant, as we demonstrated on [25, 26]. Additionally, we also deepen on the problems associated with the constant evolution of genomic data, which were identified and described in [27, 28, 11].

Second, variant interpretation guidelines have tried to standardize how variant interpretation is carried out. Here, it is important to highlight the work of the Clinical Genome Resource (ClinGen), who currently provides more specific interpretation guidelines for several clinical contexts [29, 30]. Finally, there are several tools that have attempted to reduce the variability through process automation. We are currently working on a publication that explores the difference between these tools and compares the results they produce to those of clinical experts.

– *RQ4. How to unpack the variant interpretation process?*

To unpack the variant interpretation process, it is essential to precisely define all relevant constructs within the domain. To achieve this, we have created a Unified Modeling Language (UML) class diagram [31] that: i) precisely defines the constructs involved in variant interpretation guidelines, reflecting the rule-driven nature of the process; ii) explicitly identifies the evidence used in the interpretation, representing the knowledge-driven aspect; and iii) illustrates the evaluation of a guideline against a specific DNA variant to produce its interpretation, thereby reducing variability and enhancing traceability.

More specifically, the model characterizes each interpretation guideline by its title, authors, applicability (e.g., specific clinical contexts for application), and a URL. Guidelines are composed of criteria, which can be either Boolean (evaluated as true/false) or score-based. These criteria are further broken down into specific conditions represented as metrics, which are evaluated based on evidence from various sources. The fulfillment of these metrics directly influences the outcomes of the corresponding criteria, ultimately determining the final variant interpretation. Comprehensive definitions have been established for each class in the model, including detailed descriptions of the class attributes. These definitions have been consolidated into a document that will accompany the class diagram, ensuring the model is thoroughly defined and clearly understood.

This model offers two clear benefits: (a) Definition of a common framework for representing the variant interpretation process; (b) Disentanglement of the intricate details of variant interpretation by resolving aspects whose definitions are left implicit or ambiguous, requiring clarification. The first version of this model was presented at 42nd the International Conference on Conceptual Modeling [32]. We are currently preparing an extended version of this publication for the Data & Knowledge Engineering journal.

Finally, this conceptual model has been complemented with a data model that allows for a precise and consistent representation of the evidence required for variant interpretation [33]. This data model will facilitate communication among experts, data integration and any potential automation of the interpretation process.

– *RQ5. How to operationalize the variant interpretation process?*

In its current state, variant interpretation is qualitative and unspecific by nature. We have defined a method that allows translating this impreciseness into a small and well-defined set of steps. The method involves four stages: i) selecting the guideline for interpretation, ii) defining its constructs based on the conceptual model developed in RQ4, iii) choosing the most suitable data sources to evaluate these constructs, and iv) interpreting a variant using the established framework. Breaking this complex process into more specific pieces offers the following benefits: (a) making explicit the intricate process of variant interpretation ; (b) guiding decision-making; and (c) facilitating reproducibility. This method is conceptually founded by the conceptual model described in RQ4.

6. Conclusions and Future outlook

This thesis has as main objective unpacking and operationalizing the challenging domain of variant interpretation. According to the Design Science methodology followed in this Thesis, the research's problem investigation and treatment design stages are complete. Our next steps are to validate the proposed artifacts and answer RQ6 and RQ7. The unpacking (RQ6) will be validated by carrying out a experimental evaluation where we plan to compare the textual description of the variant interpretation constructs with the definition of these constructs using the conceptual model defined in RQ4. For the operationalization process (RQ7), we are currently developing technological support for the method, which will be firmly based on the conceptual model derived from the unpacking process. This technological support is the basis for performing a Technical action report (TAR), where we will test the designed artifacts in a real context with collaboration of a clinical laboratory specialized in oncology. If possible, replications will be carried out in other clinical domains of interest.

This Ph.D. Thesis has defined strong foundations for solving an important challenge in the variant interpretation domain: the lack of precise definitions and systematization. We envision that the unpacking and operationalization proposed in this research will result in a more explainable, reproducible, and reliable interpretation process.

Acknowledgments

I want to thank Prof. Oscar Pastor (opastor@dsic.upv.es) and Dra. Ana Leon (aleon@vrain.upv.es) from the Universitat Politecnica de Valencia for the supervision of this Thesis. This Thesis is supported by the Generalitat Valenciana under the project ACIF/2021/117.

References

[1] E. Zeggini, et al., Translational genomics and precision medicine: Moving from the lab to the clinic, Science 365 (2019) 1409–1413.

[2] U.S Food and Drug Administration (FDA), Precision medicine, 2018. URL: https://www.fda.gov/medical-devices/in-vitro-diagnostics/precision-medicine.

[3] D. J. Rigden, et al., The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection, Nucleic Acids Research 51 (2023) D1–D8. doi:10.1093/nar/gkac1186.

[4] S. Harrison, et al., Overview of specifications to the acmg/amp variant interpretation guidelines, Current Protocols in Human Genetics 103 (2019). doi:10.1002/cphg.93.

[5] A. Niehaus, et al., A survey assessing adoption of the ACMG-AMP guidelines for interpreting sequence variants and identification of areas for continued improvement, Genet Med 21 (2019) 1699–1701.

[6] Y.-E. Kim, et al., Challenges and considerations in sequence variant interpretation for mendelian disorders, Annals of Laboratory Medicine 39 (2019) 421. doi:10.3343/alm.2019.39.5.421.

[7] M. Lebo, et al., Data sharing as a national quality improvement program: Reporting on brca1 and brca2 variant-interpretation comparisons through the canadian open genetics repository (cogr), Genetics in medicine : official journal of the American College of Medical Genetics 20 (2017). doi:10.1038/gim.2017.80.

[8] M. Vihinen, Problems in variation interpretation guidelines and in their implementation in computational tools, Molecular Genetics & Genomic Medicine 8 (2020) e1206. doi:https://doi.org/10.1002/mgg3.1206.

[9] N. Agaoglu, et al., Consistency of variant interpretations among bioinformaticians and clinical geneticists in hereditary cancer panels, European Journal of Human Genetics 30 (2022). doi:10.1038/s41431-022-01060-7.

[10] A. Bernasconi, et al., A comprehensive approach for the conceptual modeling of genomic data, in: J. Ralyté, S. Chakravarthy, M. Mohania, M. A. Jeusfeld, K. Karlapalem (Eds.), Conceptual Modeling, Lecture Notes in Computer Science, Springer International Publishing, 2022, pp. 194–208. doi:10.1007/978-3-031-17995-2_14.

[11] A. García S., et al., The challenge of managing the evolution of genomics data over time: a conceptual model-based approach, BMC Bioinformatics 23 (2022) 472. doi:10.1186/s12859-022-04944-z.

[12] J. F. Reyes Roman, et al., Applying conceptual modeling to better understand the human genome, in: I. Comyn-Wattiau, K. Tanaka, I.-Y. Song, S. Yamamoto, M. Saeki (Eds.), Conceptual Modeling, Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 404–412. doi:10.1007/978-3-319-46397-1_31.

[13] C. Di Ciccio, et al., Knowledge-intensive processes: Characteristics, requirements and analysis of contemporary approaches, Journal on Data Semantics 4 (2015) 29–57. doi:10.1007/s13740-014-0038-4.

[14] M. Landrum, et al., ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Research 46 (2017) D1062–D1067. doi:10.1093/nar/gkx1153.

[15] I. F. A. C. Fokkema, et al., Lovd v.2.0: the next generation in gene variant databases, Human Mutation 32 (2011) 557–563. doi:https://doi.org/10.1002/humu.21438.

[16] A. Palacio, O. Pastor, Smart data for genomic information systems: the sile method, Complex Systems Informatics and Modeling Quarterly (2018) 1–23. doi:10.7250/csimq.2018-17.01.

[17] P. J. H. Engel, Tacit knowledge and visual expertise in medical diagnostic reasoning: Implications for medical education, Medical Teacher 30 (2008) e184–e188. doi:10.1080/01421590802144260.

[18] A. Furqan, et al., Care in specialized centers and data sharing increase agreement in hypertrophic cardiomyopathy genetic test interpretation, Circulation: Cardiovascular Genetics 10 (2017) e001700.

[19] S. Richards, et al., Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology, Genetics in medicine 17 (2015) 405–423.

[20] A. Ramdaney, et al., Beware the laboratory report: Discrepancy in variant classification on reproductive carrier screening, Genetics in Medicine 20 (2017) 374–375.

[21] P. Gao, et al., Challenges of providing concordant interpretation of somatic variants in non-small cell lung cancer: A multicenter study, J Cancer 10 (2019) 1814–1824.

[22] C. Kopanos, et al., VarSome: the human genomic variant search engine, Bioinformatics 35 (2018) 1978–1980.

[23] Q. Li, K. Wang, InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines, The American Journal of Human Genetics 100 (2017) 267–280.

[24] R. J. Wieringa, Design Science Methodology for Information Systems and Software Engineering, Springer Berlin Heidelberg, 2014. doi:https://doi.org/10.1007/978-3-662-43839-8.

[25] M. Costa, et al., A comparative analysis of the completeness and concordance of data sources with cancer-associated information, in: R. S. S. Guizzardi, B. Neumayr (Eds.), Advances in Conceptual Modeling - ER 2022 Workshops, CMLS, EmpER, and JUSMOD, Hyderabad, India, October 17-20, 2022, Proceedings, volume 13650 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 35–44. doi:10.1007/978-3-031-22036-4_4.

[26] M. Costa, et al., The consequences of data dispersion in genomics: a comparative analysis of data sources for precision medicine, BMC Medical Informatics Decis. Mak. 23 (2023) 256. doi:10.1186/S12911-023-02342-W.

[27] M. Costa, et al., The importance of the temporal dimension in identifying relevant genomic variants: A case study, in: G. Grossmann, S. Ram (Eds.), Advances in Conceptual Modeling - ER 2020 Workshops CMAI, CMLS, CMOMM4FAIR, CoMoNoS, EmpER, Vienna, Austria, November 3-6, 2020, Proceedings, volume 12584 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 51–60. doi:10.1007/978-3-030-65847-2_5.

[28] A. G. Simón, et al., Characterization and treatment of the temporal dimension of genomic variations: A conceptual model-based approach, in: I. Reinhartz-Berger, S. W. Sadiq (Eds.), Advances in Conceptual Modeling - ER 2021 Workshops CoMoNoS, EmpER, CMLS, St. John's, NL, Canada, October 18-21, 2021, Proceedings, volume 13012 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 104–113. doi:10.1007/978-3-030-88358-4_9.

[29] J. Goldstein, et al., Variant classification for pompe disease; acmg/amp specifications from the clingen lysosomal diseases variant curation expert panel, Molecular Genetics and Metabolism 140 (2023) 107715. doi:10.1016/j.ymgme.2023.107715.

[30] L. Biesecker, et al., Clingen guidance for use of the pp1/bs4 co-segregation and pp4 phenotype specificity criteria for sequence variant pathogenicity classification, The American Journal of Human Genetics 111 (2023). doi:10.1016/j.ajhg.2023.11.009.

[31] G. Booch, et al., The unified modeling language, Unix Review 14 (1996) 5.

[32] M. Costa, et al., A reference meta-model to understand DNA variant interpretation guidelines, in: J. P. A. Almeida, J. Borbinha, G. Guizzardi, S. Link, J. Zdravkovic (Eds.), Conceptual Modeling - 42nd International Conference, ER 2023, Lisbon, Portugal, November 6-9, 2023, Proceedings, volume 14320 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 375–393. doi:10.1007/978-3-031-47262-6_20.

[33] M. Costa, et al., Comprehensive representation of variation interpretation data via conceptual modeling, in: T. P. Sales, J. Araújo, J. Borbinha, G. Guizzardi (Eds.), Advances in Conceptual Modeling - ER 2023 Workshops, CMLS, CMOMM4FAIR, EmpER, JUSMOD, OntoCom, QUAMES, and SmartFood, Lisbon, Portugal, November 6-9, 2023, Proceedings, volume 14319 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 25–34. doi:10.1007/978-3-031-47112-4_3.