

Towards a Complex View of Disease: A Conceptual Modeling Perspective

Diana Martínez-Minguet^{1,*}, Mireia Costa¹, Alberto García¹ and Óscar Pastor¹

¹PROS Research Group, Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camí de Vera s/n, Valencia, 46022, Spain

Abstract

Modern healthcare is shifting toward a more personalized approach, where treatments and diagnostics are tailored to the individual. Genetic testing is a key driver of this evolution, offering a powerful way to diagnose and assess health risks based on a person's unique genetic makeup. Traditionally, this analysis has focused on identifying a single, high-impact genetic variant as the primary cause of a patient's symptoms. However, advancements over the last few years have revealed that most diseases are far more complex, arising from the combined influence of numerous variants across the entire genome. This new perspective drives the daily generation of vast and complex data, creating an urgent need for genomic information systems to evolve in support of this new disease paradigm. The PROS Research Group specializes in creating genomic information systems built upon a strong conceptual modeling foundation. The cornerstone of these systems is the Conceptual Schema of the Human Genome (CSHG), which provides a standardized model for genomic data. However, this model remains grounded in the traditional, single-variant perspective. This paper presents an extension to the CSHG that integrates both single-variant and the more complex multi-variant perspective. Ultimately, this contribution provides the foundation for generating trustworthy and transparent information systems that can accurately reflect the full complexity of human disease.

Keywords

Conceptual Modeling, Conceptual Schema of the Human Genome, Complex Disease, DNA Variant

1. Introduction

The modern era of medicine is defined by the pursuit of precision: tailoring diagnostics and treatments to the unique profile of each patient [1]. Because our DNA is our most characteristic individual feature, understanding how this genetic blueprint contributes to disease lies at the heart of this effort. In the clinical setting, this knowledge is put into practice through genetic testing, which analyzes a patient's DNA to identify relevant genetic variants.

Traditionally, the goal of genetic testing has been to pinpoint a single, high-impact variant thought to be responsible for a patient's condition [2]. However, this approach is insufficient to explain the genetic architecture of most common diseases. Indeed, this single-variant perspective, often referred to as a monogenic cause of disease, is primarily applicable to two scenarios: rare disorders (e.g., sickle cell disease or Huntington's disease) which collectively affect up to 7% of the population [3], and the small fraction of cases within common diseases that are driven by a single-gene cause. Inherited cancers are a prime example, where well-known monogenic forms, such as those caused by high-impact variants in the BRCA1 or BRCA2 genes, represent only 5–10% of all cases [4].

In contrast, the genetic basis of most common diseases, clinically known as complex diseases, is far more intricate. Rather than being caused by a single variant, these conditions arise when a particular set of variants acts collectively to increase an individual's risk—or predisposition—to developing the

ER2025: Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos, October 20-23, 2025, Poitiers, France

*Corresponding author.

✉ dmarmin@vrain.upv.es (D. Martínez-Minguet); micossan@vrain.upv.es (M. Costa); algarsi3@vrain.upv.es (A. García); opastor@dsic.upv.es (Ó. Pastor)

ORCID 0009-0002-3191-1969 (D. Martínez-Minguet); 0000-0002-8614-0914 (M. Costa); 0000-0001-5910-4363 (A. García); 0000-0002-1320-8471 (Ó. Pastor)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

disease [5]. This represents a fundamental shift in understanding, moving from a *monogenic* model of direct causation to a *polygenic* model based on the cumulative effect of numerous variants.

Therefore, the future of clinical genetics depends on incorporating this polygenic perspective into standard testing. This future is materializing through Polygenic Risk Scores (PRSs), which consolidate an individual's complex genetic information into a single, actionable metric [6]. A PRS is calculated by a statistical model that aggregates the small effects of thousands, or even millions, of common genetic variants from across the genome. The result is a single number that quantifies an individual's inherited predisposition to a specific disease.

While this measure is not definitive, integrating such polygenic information into clinical practice has the potential to enable earlier and more precise disease prevention strategies, tailor screening protocols to an individual's genetic profile, and guide lifestyle or therapeutic interventions more effectively [7, 8]. Furthermore, by moving beyond the single-variant paradigm, genetic testing will shift from a primarily diagnostic tool for rare conditions and the small percentage of common conditions caused by a single variant to a proactive instrument for managing health across the broader population.

However, to achieve this transformation, the genomic information systems that support genetic testing must evolve to manage and interpret the increased complexity of polygenic data. In addition, multidisciplinary teams involved (genetic counselors, bioinformaticians, and technical experts) will need to adapt to new and rapidly evolving knowledge. In this context, conceptual modeling provides crucial support, offering a way to represent domain-specific information that is intuitive, easy to understand, and meaningful [9]. In this context, the PROS Research Group at the Polytechnic University of Valencia has established a strong foundation with its Model-based Development of trustworthy and transparent genomic information systems. At the core of these systems lies the Conceptual Schema of the Human Genome (CSHG) [10], a framework designed to represent everything from the genome itself to the way genetic variants influence disease.

As polygenic approaches gain importance in clinical settings, this work presents a natural evolution of the CSHG to integrate this perspective of polygenic diseases while preserving its proven monogenic capabilities. Our main contribution is a unified model that considers both monogenic and polygenic disease paradigms. This evolution will enable the development of the next generation of model-based information systems to capture the full spectrum of human genetic disease, allowing for more comprehensive and clinically relevant applications.

The remainder of this paper is structured as follows: Section 2 reviews the key genetic concepts underlying our model extension. Section 3 then presents our proposed extension to the Conceptual Schema of the Human Genome (CSHG). Finally, Section 4 summarizes our contributions and discusses future research directions.

2. Background

To provide context for our work, this section presents essential background on the modern understanding of genetics in disease. First, we will cover the role of genetics in complex diseases. Then, we will explain Polygenic Risk Scores (PRSs), the modern method used to measure the genetic risk for these conditions.

2.1. Genetics in Complex Disease

Complex diseases are conditions whose development is the result of the combined action of multiple genes and environmental factors, rather than a single, high-impact genetic variant [5]. Most common health conditions, including cardiovascular disease, type 2 diabetes, and many cancers, fall into this category.

At the genetic level, these diseases are shaped by the cumulative effect of thousands of common DNA variants. To identify the variants that influence disease risk, researchers use Genome-Wide Association Studies (GWAS) [11]. A GWAS effectively scans the entire genome of tens of thousands of individuals and compares their DNA profiles to search for variants that occur more frequently in people with a given condition than in those without it.

While any single variant discovered through GWAS typically has only a small effect, these studies have proven that the cumulative impact of these variants is significant. This insight—that many small-effect variants together can explain a large part of an individual’s disease susceptibility—led directly to the development of statistical models called Polygenic Risk Score (PRS) models [12]. These models aggregate the effects of numerous variants to quantify a person’s inherited predisposition to a specific disease. The output is a single, comprehensive number which itself is also referred to as the individual’s PRS.

2.2. Polygenic Risk Scores and their Clinical Potential

The output of a Genome-Wide Association Study (GWAS) is a list of candidate variants statistically associated with a given disease. However, these raw results are not a direct measure of an individual’s overall risk. Instead, they serve as a starting point, a map highlighting regions of the genome that warrant further investigation to understand their true contribution to disease.

The goal of a PRS model is to transform the raw data from a GWAS into a single, clinically meaningful measure of an individual’s genetic risk [12, 13]. To do so, the list of variants identified by GWAS is curated. Statistical methods are used to filter out “noisy” or low-quality variants and to adjust for variants that only appear to be associated with the disease due to their proximity to a truly causative variant. As a result, a refined variant set is established, where each variant is assigned a “risk weight” (its effect size) reflecting its specific contribution to the disease.

This curated set of variants and their corresponding weights forms the core of the PRS model itself, ready to be applied to an individual’s genetic data. An individual’s PRS is calculated by summing the assigned risk weights of all the variants from this set that are present in the person’s DNA, resulting in a single number that represents their overall inherited predisposition.

The clinical interpretation of this individual’s PRS depends on two performance metrics of the variant set that composes the PRS model: how much (risk association metrics) and how well (discriminatory power metrics) the variant set predicts a trait or disease [14, 12]. A risk association metric describes how the variant set relates to the likelihood of developing the disease. This is often expressed using metrics like the Odds Ratio (OR), which compares the risk in a high-scoring group (e.g., the top 10%) to a reference group (e.g., the middle 40-60%). For instance, if a PRS model for heart disease has an OR of 2 for individuals in the top decile vs. the middle quintile, it means those in the top range are about twice as likely to develop the condition compared to those with average scores.

The other relevant performance metric is the discriminatory power, which is a measure of the variant set’s ability to differentiate between individuals at risk and those not at risk. An example is the Area Under the Curve (AUC), which gives the probability that a randomly chosen case has a higher PRS than a randomly chosen control. As an example, if a PRS model has an AUC of 0.9, it means that the model classifies correctly (i.e., assigns a higher PRS to a case than a control) 90% of the time.

Using these metrics, the PRS computed for an individual can be interpreted [15]. For instance, if the resulting PRS of an individual falls within the top decile, the individual would be twice as likely to develop the condition, while if the individual falls in the middle range, there is no increased risk of disease. This affirmation would further be supported by the fact that the model has a good discriminatory ability of 90%.

3. Evolution of the Conceptual Schema of the Human Genome

At the PROS Research Group, genomic information systems are developed to facilitate the management of genetic data for both clinical and research purposes. These systems are grounded in, and ontologically supported by, the Conceptual Schema of the Human Genome (CSHG). This ontological foundation enables the systems to be readily extended with new functionalities and adapted to evolving requirements or changes in the genomic domain.

3.1. Current state of the CSHG: single-variant approach

In the current state of the CSHG (see Fig.1), a genetic Variant is related to a Phenotype —a term that encompasses any observable trait, such as eye color, or a specific disease— through a Significance. This Significance represents the clinical interpretation a specific Variant has for a given Phenotype, as provided by the scientific community. It is characterized by two relevant attributes: a “ClinicalSignificance” and a “levelOfCertainty”.

On the one hand, the “ClinicalSignificance” describes the variant’s impact on the manifestation of the phenotype. Standard significances include Pathogenic, for variants known to cause the disease; Benign, for those not believed to have a causal role; and Uncertain Significance (VUS), where there is insufficient or conflicting evidence to determine the variant’s effect. On the other hand, the “levelOfCertainty” represents the level of confidence in the assigned “ClinicalSignificance”. This certainty is directly dependent on the quality and strength of the scientific evidence used for the assessment.

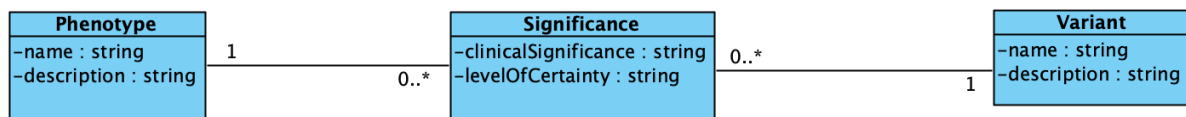


Figure 1: Model excerpt for single-variant analysis in genetic disease in the CSHG. This represents the previous approach, prior to the work presented herein.

It is important to note that a given variant may hold different Significances for different phenotypes. Even for the same phenotype, this Significance can vary depending on the expert opinion or the study from which the assessment originates. How these conflicting interpretations are managed by the CSHG is explained in [16].

While this representation supports clear and detailed mapping between individual variants and phenotypes, it cannot capture genetic predisposition to complex diseases, where risk arises from the combined influence of many variants. Below, we describe how we have addressed this gap through an extension of the CSHG.

3.2. Extension of the CSHG: multi-variant approach

The limitations of the CSHG’s single-variant representation highlight the need for an extension capable of capturing the polygenic nature of complex diseases. To address this, we have incorporated the multi-variant perspective into the model. The additions to the CSHG that materialize this new perspective are depicted in green in Figure 2.

Our modeling strategy for this extension intends to mirror the structure of the single-variant perspective. First, we shifted the focus from individual Variants to VariantSets. This approach allows us to evaluate multiple Variants jointly, which is essential for representing the genetic basis of complex diseases. In the VariantSet, each of the Variants is assigned an EffectSize sets that represents its contribution to the disease risk. A single Variant may belong to multiple variant sets, with a distinct effect size for each specific set.

Each VariantSet comes originally from a GenomeWideAssociationStudy (GWAS). It is important to clarify that a VariantSet represents the curated result after statistical processing (see Section 2.2), not the raw GWAS output. However, this link to the original GWAS, though indirect, is explicitly maintained in the model to ensure full data traceability. Furthermore, because different curation processes can be applied to the same initial data, a single GenomeWideAssociationStudy can yield multiple, distinct VariantSets. This one-to-many relationship is explicitly defined by the cardinalities between these two classes in our model. The specific curation process and other relevant details for how a VariantSet was obtained are captured in its “name”, “statisticalMethod” and “description” attributes.

Second, with the focus now on VariantSets, we need to extend the analogy from the single-variant model to represent their collective significance for a Phenotype. As detailed in Section 2.2, the clinical

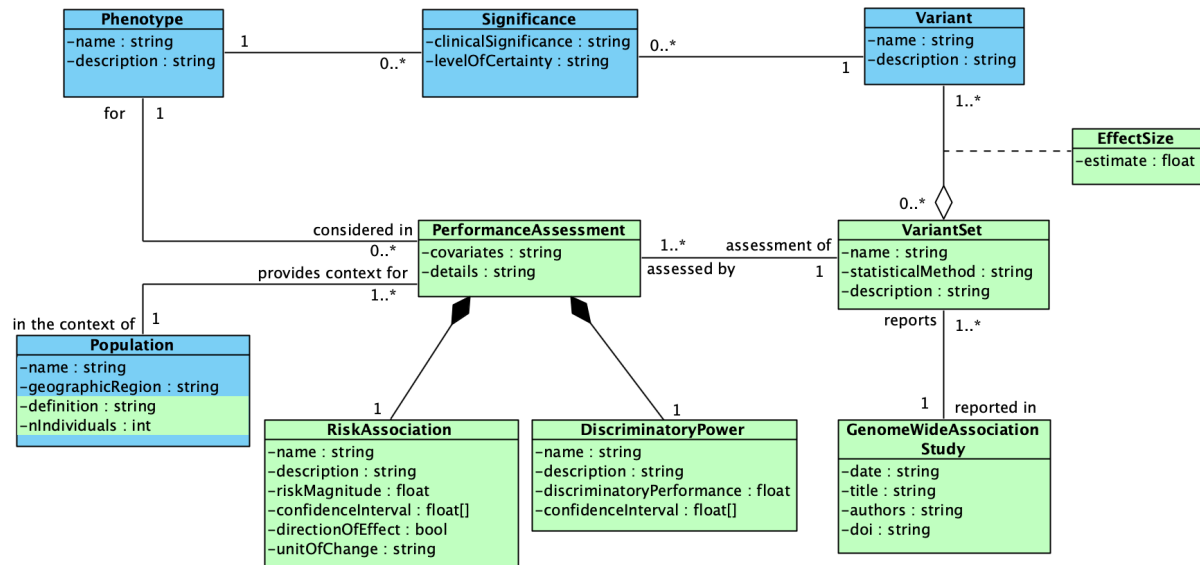


Figure 2: Model excerpt for the new proposed multi-variant analysis in genetic disease in the CSHG.

interpretation of a PRS is guided by two performance metrics: risk association (measuring how much risk the variant set confers) and discriminatory power (assessing how well it predicts the disease). These metrics create a direct analogy to the single-variant model. A set’s RiskAssociation, which quantifies the statistical link to the phenotype, is analogous to “ClinicalSignificance”. Its DiscriminatoryPower, which assesses how well the set distinguishes cases from controls, is analogous to “levelOfCertainty”. This relationship between the single-variant and multi-variant approaches is summarized in Table 1.

Table 1
An analogy between single-variant and multi-variant approaches.

Single-variant approach	Definition	Multi-variant approach	Definition
Clinical significance , indicates qualitative measures. <i>Example:</i> pathogenic or benign.	Strength of evidence that a variant is linked to a disease/phenotype. <i>Example:</i> pathogenic (i.e., disorder causing).	Risk association metrics. <i>Example:</i> Odds Ratio (OR), Hazard Ratio (HR), <i>beta</i> coefficient, etc.	Quantify the statistical association between the aggregated variant set and the predicted phenotype. <i>Example:</i> $OR = 2$, i.e., individuals with a higher PRS have about twice the chance of developing the disease compared to those with an average PRS.
Level of certainty , indicates qualitative descriptors. <i>Example:</i> high, moderate, or low evidence.	Confidence in the classification, based on strength, consistency, and quality of evidence. <i>Example:</i> moderate evidence.	Discriminatory power metrics. <i>Example:</i> AUC (Area Under ROC Curve), C-index (Concordance index), etc.	Measure of the variant set’s ability to differentiate between individuals at risk and not at risk. <i>Example:</i> $AUC = 0.9$, i.e., the probability that a randomly chosen case has a higher PRS than a randomly chosen control is 90%.

Elaborating on these metrics, the RiskAssociation is defined by a “name” and a “description” that characterize its specific implementation (e.g., as an Odds Ratio (OR), Hazard Ratio (HR), or β coefficient). The actual value of the metric is represented by the “riskMagnitude”, and being a statistical measure, it is accompanied by its corresponding “confidenceInterval”. The interpretation of the “riskMagnitude” requires considering two factors: the “unitOfChange” and the “directionOfEffect”. The “unitOfChange” represents the scale of comparison, specifying exactly which groups are being compared. For example, a PRS model might compare individuals in the top decile (the top 10% of scores) against those in

the middle quintile (the middle 20% of scores). If the resulting Odds Ratio (OR) is 2.0, it specifically means that people in that top 10% bracket have twice the odds of developing the disease compared to those in the middle 20% reference group. Without defining these comparison groups, the number itself is meaningless. The “directionOfEffect” represents whether the calculated “riskMagnitude” is risk-conferring– the typical situation– or represents a protective effect. For instance, in the case of the odds ratio type of RiskAssociation, an $OR > 1$ represents the association is risk-conferring, while $OR < 1$ represents a protective effect.

Similarly, the DiscriminatoryPower metric is defined by a “name” and “description” specifying the metric used (e.g., AUC). The resulting value is stored in the “discriminatoryPerformance” attribute and, like the “riskMagnitude”, is accompanied by its “confidenceInterval”.

These two metrics are encapsulated within a PerformanceAssessment class. This class completes the analogy with the single-variant perspective by linking a VariantSet to a Phenotype, using the RiskAssociation and DiscriminatoryPower as its fundamental components. The PerformanceAssessment also includes “covariates” (such as age or sex) that may influence the results, along with any other important “details”.

A performance assessment’s validity is strictly tied to a specific Population, making this context as fundamental as the Phenotype for correct interpretation. For example, metrics derived from a study of middle-aged European women cannot be reliably applied to an elderly Asian man, given their vastly different genetic and environmental backgrounds. To model this, we extended the existing Population class from the original CSHG. While it was previously defined by a “name” and “genomicRegion”, our extension adds a “definition” attribute (to specify demographics like age or sex) and the number of individuals (“nIndividuals”) in the study cohort.

With this representation, we can precisely define complex statements such as the one introduced in Section 2.2. For instance, consider the following example from a type 2 diabetes variant set developed in [17] (PGS Catalog ID PGS001781), and depicted in Figure 3: “A type 2 diabetes mellitus variant set has an OR of 1.75 per standard deviation (SD) for European individuals aged 57 years. This indicates that individuals whose PRS falls at 1 SD when compared to a distribution of these population characteristics have an increased risk of developing the disease 1.75 times more than people with an average score. In addition, the variant set can correctly classify individuals with and without type 2 diabetes mellitus 73% of the time”. As depicted in the figure, the variant set named T2D_PRSCS has been obtained through the PRS-CS-auto statistical method, we also added two variants with their respective effect sizes as an example, although the total number is 1091673 aggregated variants.

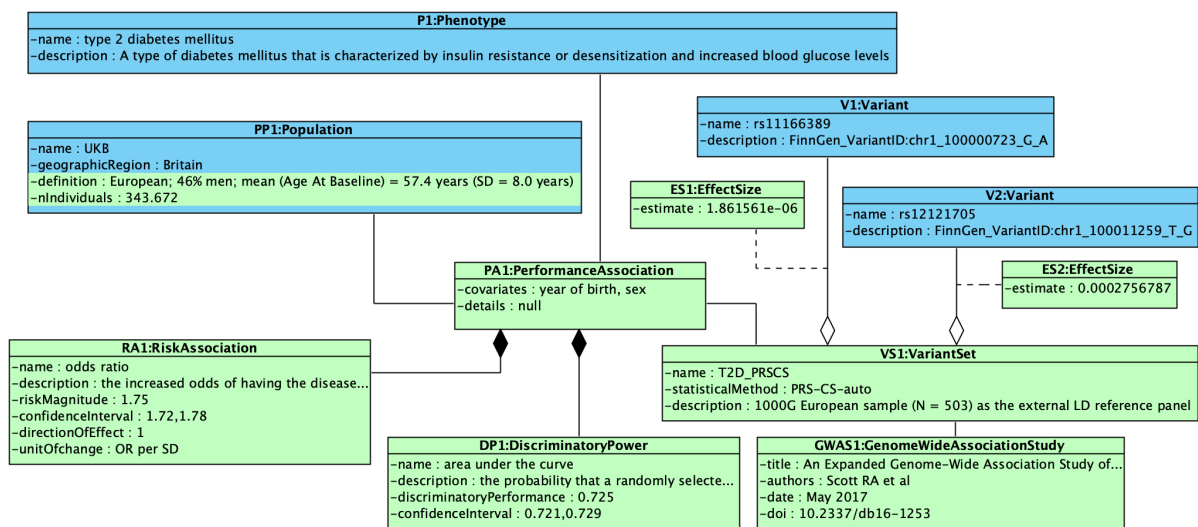


Figure 3: Object diagram exemplifying the changes in the CSHG to work with a multi-variant approach. This example shows a variant set associated with type 2 diabetes mellitus.

This shift from a single-variant to a multi-variant perspective is essential for integrating polygenic

measures, such as Polygenic Risk Scores, into genomic information systems and for enabling their clinical use.

4. Conclusion and Future work

Genetic testing offers major benefits for the precision medicine context, allowing for targeted interventions and precise health risk assessments. While the traditional monogenic paradigm of disease is routinely adopted, the majority of common diseases do not conform to this pattern. This work addresses a pressing gap in genomic information systems, which call for the adoption of the polygenic nature of common complex diseases.

We have introduced an extension to the Conceptual Schema of the Human Genome that unifies monogenic (single-variant approach) and polygenic (multi-variant approach) perspectives on genetic disease. The multi-variant approach is proposed as an analogy to the current single-variant approach, thus facilitating its understanding and uptake through a known parallelism. With this innovation, we enable the representation of the complexity of modern disease genetics within a coherent, model-driven framework.

Additionally, thanks to the standardized representation of information provided by the schema, we can support and promote research through a variety of tasks. For instance, it is well established that the same phenotype can be associated with multiple variant sets, depending on the initial GWAS data and the statistical post-processing applied. The standardized representation enables straightforward comparison of performance assessments across different variant sets, helping to determine which set is most suitable for a given genomic analysis.

At the variant level, the schema allows the study of intersections between different variant sets, facilitating the identification of overlapping variants and the comparison of their effect sizes. Comparing variant sets for the same phenotype can provide insights on their concordance, while comparing sets across different phenotypes can uncover correlations between comorbid diseases, as seen in certain mental disorders. Additionally, the unified representation supports analyses such as determining whether a single high-impact variant linked to one phenotype may also contribute to another phenotype with a smaller effect size.

Regarding the future schema itself, next steps include refining the schema's semantics through validation with experts' feedback, testing it against concrete and validated PRS models that aim to be included as part of genomic analyses for polygenic risk prediction, and integrating it into operational genomic platforms as the main organizational data structure to manage PRS-related data and perform reliable interpretations for polygenic risk predictions. This unified representation opens the door to comprehensive analyses, such as evaluating the implications of an individual carrying both a high-impact variant and a high polygenic risk for the same phenotype, or conversely, exhibiting complementary genetic risks arising from a benign high-impact variant but a high risk from common variation.

The advancement proposed in this study is fundamental because current genetic testing workflows do not yet incorporate PRS information. By providing a clear and transparent way to represent PRS-related data, our proposed extension lays the essential groundwork for future genetic information systems incorporating PRS information. Moreover, because the model was derived through a deliberate parallel with the well-established single-variant approach, it lowers barriers to understanding and adoption, ensuring a smoother transition to the polygenic perspective in both research and clinical practice. By combining rigorous conceptual modeling with emerging genomic needs, this work paves the way for information systems that more faithfully reflect the true complexity of human disease.

Acknowledgments

This work was supported by the Generalitat Valenciana through the CoMoDiD project (CIPROM/2021/023) and the predoctoral grant (ACIF/2021/117), and by the Universitat Politècnica de València through grant PAID-06-24.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] J. Denny, F. Collins, Precision medicine in 2030—seven ways to transform healthcare, *Cell* 184 (2021) 1415–1419. doi:10.1016/j.cell.2021.01.015.
- [2] N. Franceschini, A. Frick, J. Kopp, Genetic testing in clinical settings, *American Journal of Kidney Diseases* 72 (2018). doi:10.1053/j.ajkd.2018.02.351.
- [3] S. Wakap, D. Lambert, A. Olry, C. Rodwell, C. Gueydan, L. Valérie, D. Murphy, Y. Cam, A. Rath, Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database, *European Journal of Human Genetics* 28 (2019). doi:10.1038/s41431-019-0508-0.
- [4] S. Hodgson, W. Foulkes, E. Maher, C. Turnbull, Inherited susceptibility to cancer: Past, present and future, *Annals of Human Genetics* 89 (2025). doi:10.1111/ahg.70013.
- [5] P. Visscher, et al., Discovery and implications of polygenicity of common diseases, *Science* 373 (2021) 1468–1473. doi:10.1126/science.abi8206.
- [6] Y. Ma, et al., Genetic prediction of complex traits with polygenic scores: a statistical review, *Trends in Genetics* 37 (2021). doi:10.1016/j.tig.2021.06.004.
- [7] N. Wray, T. Lin, J. Austin, J. McGrath, I. Hickie, G. Murray, P. Visscher, From basic science to clinical application of polygenic risk scores: A primer, *JAMA psychiatry* 78 (2020). doi:10.1001/jamapsychiatry.2020.3049.
- [8] C. Lewis, E. Vassos, Polygenic risk scores: From research tools to clinical instruments, *Genome Medicine* 12 (2020). doi:10.1186/s13073-020-00742-5.
- [9] N. Guarino, G. Guizzardi, J. Mylopoulos, On the philosophical foundations of conceptual models, in: *Information Modelling and Knowledge Bases XXXI, Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 1–15. doi:10.3233/FAIA200002.
- [10] A. García S., A. Palacio, J. Reyes Román, J. Casamayor, O. Pastor, A conceptual model-based approach to improve the representation and management of omics data in precision medicine, *IEEE Access PP* (2021) 1–1. doi:10.1109/ACCESS.2021.3128757.
- [11] E. Uffelmann, Q. Huang, N. S. Munung, J. Vries, Y. Okada, A. Martin, H. Martin, T. Lappalainen, D. Posthuma, Genome-wide association studies, *Nature Reviews Methods Primers* 1 (2021). doi:10.1038/s43586-021-00056-9.
- [12] C. Babb de Villiers, et al., Understanding polygenic models, their development and the potential application of polygenic scores in healthcare, *Journal of Medical Genetics* 57 (2020). doi:10.1136/jmedgenet-2019-106763.
- [13] R. Xiang, M. Kelemen, Y. Xu, L. Harris, H. Parkinson, M. Inouye, S. Lambert, Recent advances in polygenic scores: translation, equitability, methods and fair tools, *Genome Medicine* 16 (2024). doi:10.1186/s13073-024-01304-9.
- [14] A. Lewis, R. Green, J. Vassy, Polygenic risk scores in the clinic: Translating risk into action, *Human Genetics and Genomics Advances* 2 (2021) 100047. doi:10.1016/j.xhgg.2021.100047.
- [15] M. Corpas, K. Megy, A. Metastasio, E. Lehmann, Implementation of individualised polygenic risk score analysis: a test case of a family of four, *BMC Medical Genomics* 15 (2022). doi:10.1186/s12920-022-01331-8.
- [16] A. Palacio, A. S., M. Costa, A. Ribelles, O. Pastor, Evolution of an Adaptive Information System for Precision Medicine, 2021, pp. 3–10. doi:10.1007/978-3-030-79108-7_1.
- [17] M. Tamlander, N. Mars, M. Pirinen, A. Palotie, M. Daly, B. Riley-Gillis, H. Jacob, D. Paul, H. Runz, S. John, R. Plenge, J. Maranville, G. Okafo, N. Lawless, H. Salminen-Mankonen, M. McCarthy, J. Hunkapiller, M. Ehm, K. Auro, T. Southerington, Integration of questionnaire-based risk factors

improves polygenic risk scores for human coronary heart disease and type 2 diabetes, *Communications Biology* 5 (2022) 158. doi:10.1038/s42003-021-02996-0.