Translating Polygenic Risk Score Research to a Clinical Setting

Diana Martínez-Minguet[1,*]

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

Abstract

Polygenic Risk Score (PRS) research represents an emerging and active area of medical genetics. A PRS allows to estimate the extent to which an individual's genetics contributes to the development of a complex disease. PRS analyses can be useful in many research areas, allowing personalized treatment and risk stratification of the population. To conduct a PRS analysis, a PRS Model is required. Clinicians face the critical task of selecting the most accurate PRS Model for their analysis. This is a crucial step given the impact on effectiveness of the analysis results, which can compromise patient care. However, variability of PRS Models, heterogeneity in concept representation and complexities regarding the prioritization of PRS Models in terms of relevance, make PRS Model selection a demanding task. This prevents the effective application of PRS analyses outside of its area of expertise, hindering their translation to the clinical setting. Following the Design Science methodology, we propose the design and validation of two artifacts to overcome these barriers: (i) a conceptual model for easing domain comprehension and PRS Models comparison and (ii) a method to allow for an adequate prioritization of PRS Models. This thesis aims to streamline the PRS Model selection process, assisting in the PRS analysis application and thereby helping to bridge the gap between PRS research and clinical practice.

Keywords

Polygenic Risk Score, Conceptual Modeling, Precision Medicine

1. Introduction

Precision medicine has revolutionized how diseases are diagnosed, treated and prevented [1]. An individual's genomics, environmental, and lifestyle factors are now considered for providing personalized care. Genomics plays an emerging role in medical research, providing insights in how the human genome contributes to disease [2]. One can differentiate between *simple* and *complex* diseases. While simple diseases have a clear genetic cause that can be traced back to a single change in our DNA (i.e., genetic variant), complex diseases are caused by a combination of several genetic variants, each contributing differently to the risk of developing such a disease [3]. Research on the genetic causes of complex diseases is of great interest since these represent the majority of common diseases, posing the greatest burden on health care.

In recent years, **Polygenic Risk Scores** have emerged as a new tool to estimate to what extent the genetics of an individual contributes to a complex disease [4]. In fact, genetic risk analyses using PRSs –**PRS analyses**– are one the most promising approaches for improving

CAiSE 2024 Doctoral Consortium

*Corresponding author.

dmarmin@vrain.upv.es (D. Martínez-Minguet)

0009-0002-3191-1969 (D. Martínez-Minguet)

clinical decision-making assistance, treatment choices, and risk stratification of the population for complex diseases.

To conduct a PRS analysis a **PRS Model** is required. A PRS Model applies statistical methods and algorithms to genomic population data in order to identify the variants associated with the complex disease under study [5]. PRS Models can be developed in a variety of ways, varying in the ancestry of the population data or the statistical method used, among other factors. This results in a wide range of available PRS Models, each with its unique characteristics [6], that need to be compared in order to find the most suitable one for a given PRS analysis.

Clinicians face the critical task of selecting the most accurate PRS Model for their analysis, which is a crucial step given that an inadequate model can compromise the effectiveness of the analysis results [6]. However, the selection process is hindered by complexities of the domain, regarding lack of standardization in PRS Model reporting and variability of PRS Models [7, 8]. There are recommendations in the literature that can be leveraged to face the stated problems and aid in the PRS Model selection process, but there is no methodology that formally addresses this issue in a systematized way. For this reason, we propose the application of information systems engineering techniques in order to accurately systematize the PRS Model selection process.

To sum up, this PhD thesis aims to streamline the PRS Model selection process, ensuring that clinicians can choose the best PRS Model for their analysis. Following the Design Science methodology [9], we will tackle the challenges by the design and validation of two artifacts: (i) a Conceptual Model for characterizing the PRS research domain, providing domain clarification and easing the comparison between PRS Models and (ii) a Method for the prioritization of PRS Models, allowing for the ordering of PRS Models in terms of suitability for a given PRS analysis.

The remaining of the paper is organized as follows: Section 2 presents the problem statement and existing solutions, Section 3 defines the objectives and research questions, Section 4 yields the methodology to be followed and the expected artifacts, Section 5 oversees the current status of the research, and in Section 6 we explain the contributions of this work and draw the main conclusions.

2. Problem Statement and Existing Solutions

PRS analyses are used for estimating the genetic risk of developing a complex disease, and they represent a useful tool for clinicians, epidemiologists and other researchers. Our stakeholders are clinicians and researchers willing to perform a PRS analysis and their goal is to select the most suitable PRS Model, since the effectiveness of the analysis results will depend on the accurate selection of the PRS Model.

There are two main problems preventing the stakeholders to achieve their goal. The first issue is the heterogeneity in concept representation that hampers domain comprehension and PRS Model benchmarking. There is a lack of standardized reporting criteria and terminology not only in studies reporting PRS Models [7] but also in studies providing explanations on the PRS research domain [6, 10, 8]. As a result, the first step of the selection process which is **to compare different PRS Models becomes a demanding task**. The second issue is related to the factors that need to be taken into account when selecting an accurate PRS Model. There

are many features to be considered, for instance the ancestry distribution of the model, the statistical method used or the performance metrics evaluated [6]. These features do not provide a straightforward criterion for including or discarding a PRS Model; instead, they influence the suitability of a PRS Model for the desired analysis to varying degrees. Therefore, **it is challenging to order the PRS Models in terms of relevance or suitability for an analysis**, i.e., to prioritize them, complicating the selection process.

In the literature we can find guidelines for improving the reporting of PRS Models [7], and also recommendations for choosing a suitable PRS Model for an analysis [6]. However, there is no systematized or methodological approach for dealing with the PRS Model selection process, which is the gap we aim to address in this work. For this purpose, information systems engineering techniques can provide the means to tackle the aforementioned issues and streamline the PRS Model selection process.

3. Objectives and Research Questions

The **main goal** of this thesis is to streamline the PRS Model selection process. The Objectives (O) and Research Questions (RQs) are the following:

O1. Review of the state of the art and existing barriers in translation of PRS research into the clinical context.
RQ1.1 Which barriers exist in the PRS research domain preventing their translation to a clinical setting?
RQ1.2 Which solutions exist to mitigate the identified barriers?

O2. Design artifacts to enhance understanding of the domain and assist in the PRS analysis application.
RQ2.1 How to improve domain comprehension?
RQ2.2 How to enable PRS Model prioritization?

O3. Validate the designed artifacts and analyze contributions of this work.
RQ3.1 To what extent does the proposed solution improve domain comprehension?
RQ3.2 To what extent does the proposed solution enable PRS Model prioritization?

4. Methodology and Expected Artifacts

The methodology to be followed is Design Science, proposed by Wieringa [9]. This methodology consists of the design and investigation of artifacts in a context, in order to improve some aspect of the context. In this case, the artifacts are the conceptual model and the method described above, and the context is the PRS research field. The Design Cycle consists of three stages, namely, Problem Investigation, Problem Treatment, and Treatment Validation. The objectives of the PhD thesis described in Section 3 are aligned with the three stages (see Figure 1).

4.1. Problem Investigation

In the first stage, the associated objective to be pursued is **O1. Review of the state of the art and existing barriers in translation of PRS research into the clinical context**. We will

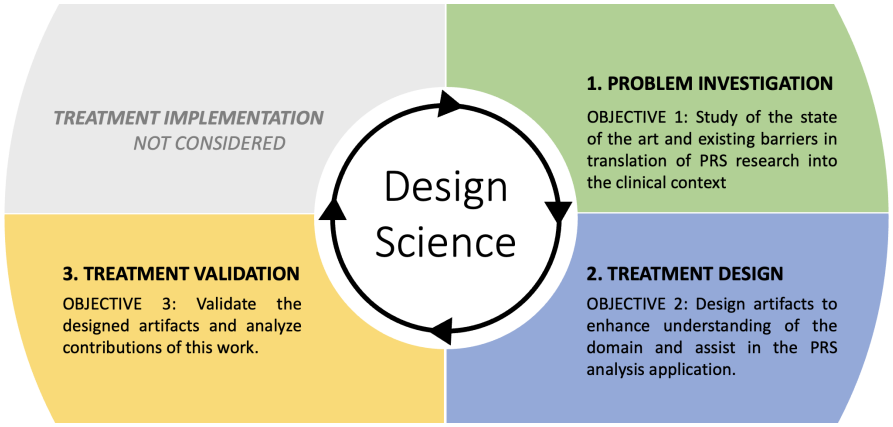

Figure 1: Design cycle of the PhD thesis.

carry out an in-depth research of the context and problem statement by answering two research questions. The first one **RQ1.1** *Which barriers exist in the PRS research domain preventing their translation to a clinical setting?* will delineate the problem to be addressed, from the viewpoints framed in the thesis research area. Here we will establish the stakeholders to be considered and their goals, which will guide the development of artifacts. In **RQ1.2** *Which solutions exist to mitigate the identified barriers?* the existing solutions found in related works will be studied in order to detect gaps or points of improvement. At this point we will define the approach to be followed to address the research problem.

4.2. Treatment Design

In this second stage we cover **O2. Design artifacts to enhance understanding of the domain and assist in the PRS analysis application**, where we will develop the expected artifacts for the solution of the identified problems.

The first foreseen artifact will result from **RQ2.1** *How to improve domain comprehension?* We will rely on the use of Conceptual Modeling (CM) techniques to provide sound knowledge on the domain of application [11]. CM has proven to be an effective solution to improve understandability and standardization in the biological domain [12, 13, 14, 15]. Therefore, the anticipated outcome is a conceptual model representing the PRS research domain, which will be created utilizing the Unified Modeling Language (UML) Class Diagram [16]. Accurate choice of nomenclature will be addressed by considering guideline recommendations and literature review [7].

Regarding the second research question, **RQ2.2** *How to enable PRS Model prioritization?*, we envisage the design of a method that will allow for an adequate ordering of PRS Models in terms of relevance with respect to a given analysis. The method will be supported by the conceptual model of the domain previously developed. The accurate characterization of the constructs defining a PRS Model will guide the definition of the requirements and considerations for developing the prioritization method. As a result, we will facilitate the prioritization of models based on criteria established by the clinician or researcher conducting the PRS analysis.

4.3. Treatment Validation

The last stage of the design cycle is the validation of the generated artifacts. We will evaluate if the stakeholders' goals are accomplished through the objective **O3. Validate the designed artifacts and analyze contributions of this work**. The two artifacts to be validated are the conceptual model of the PRS research domain and the method to enable the prioritization of PRS Models.

We envisage the artifacts validation following proposals provided by Wieringa. In order to answer **RQ3.1** *To what extent does the proposed solution improve domain comprehension?*, we forsee an empirical validation of the conceptual model with users. For the second question, **RQ3.2** *To what extent does the proposed solution enable PRS Model prioritization?* we foresee the validation of the method by expert opinion, and/or via a case study in a real setting.

5. Current status

The research is starting its second year of development. So far we have addressed **O1**, in order to answer **RQ1.1** and **RQ1.2**. The research has involved exhaustive literature review on the state of the art and the context of PRS research. As part of the context research which involves genetics of complex diseases we conducted a Systematic Mapping Review of databases with genes associated to complex diseases, identifying lack of consensus among database criteria of inclusion and classification of genes. As a result of the sate of the art on PRS research we have delimited the problem statement and the approach to be followed. This objective allowed to structure the research plan which served as guidance for this manuscript.

We have started to tackle **O2**, where artifacts are designed. Regarding **RQ2.1** we are developing a preliminary conceptual model on PRS research domain using CM techniques. In this second year we plan to keep enriching the model with additional iterations, generating new versions.

Next steps are the design of the prioritization method. First, we will define the requirements and considerations taking into account literature review and discussion with experts in the domain, i.e., which features are relevant for PRS Model selection, existing dependencies or feature priority. Secondly, we will characterize the strategy for evaluating the features. We will develop a tool support to enable the instantiation of the method and the validation of the artifacts. The validation of the artifacts are foreseen to be developed during the third year.

6. Contributions and final conclusions

The artifacts proposed in this work represent the main contributions of the research: (i) the conceptual model will provide comprehension on the PRS research domain by characterizing the relevant constructs and features thus facilitating the comparison between PRS Models, and (ii) the method to enable the prioritization of PRS Models will enable the ordering of PRS Model in terms of suitability for a given analysis. On a more general scale, this research will be contributing to the translation of PRS research into clinical practice, by streamlining the PRS Model selection process and thus aiding in the PRS analysis application.

Nowadays, medical research considers genomic, environmental and lifestyle factors of an individual for providing tailored diagnosis and treatment. Since genetic information defines an individual unequivocally, PRS analyses can be highly valuable for personalized medical care for complex diseases. Genomic research produces data, tools and new knowledge incessantly, we need to catch up to this evolution to make it profitable in real-world medical settings.

Acknowledgments

I want to thank Prof. Oscar Pastor (opastor@dsic.upv.es) and Dr. Alberto García (algarsi@vrain.upv.es) from the Universitat Politècnica de València for the supervision of this Thesis. I also want to thank PhD students Mireia Costa and René Noel for their fruitful discussions. This work was supported by the Generalitat Valenciana through the CoMoDiD project (CIPROM/2021/023).

References

[1] G. Gameiro, et al., Precision medicine: Changing the way we think about healthcare, Clinics 73 (2018). doi:10.6061/clinics/2017/e723.

[2] V. Pattan, et al., Genomics in medicine: A new era in medicine, World Journal of Methodology (2021) 2021. doi:10.5662/wjm.v11.i5.231.

[3] P. Visscher, et al., Discovery and implications of polygenicity of common diseases, Science 373 (2021) 1468–1473. doi:10.1126/science.abi8206.

[4] N. Wray, et al., From basic science to clinical application of polygenic risk scores: A primer, JAMA psychiatry 78 (2020). doi:10.1001/jamapsychiatry.2020.3049.

[5] 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.

[6] J. Collister, et al., Calculating Polygenic Risk Scores (PRS) in UK Biobank: A practical guide for epidemiologists, Frontiers in Genetics 13 (2022). doi:10.3389/fgene.2022.818574.

[7] H. Wand, et al., Improving reporting standards for polygenic scores in risk prediction studies, Nature 591 (2021) 211–219. doi:10.1038/s41586-021-03243-6.

[8] S. W. S. Choi, et al., Tutorial: a guide to performing polygenic risk score analyses, Nature Protocols 15 (2020). doi:10.1038/s41596-020-0353-1.

[9] R. J. Wieringa, Design science methodology: For information systems and software engineering, Springer Berlin Heidelberg, 2014. doi:10.1007/978-3-662-43839-8.

[10] 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.

[11] A. Olivé, Conceptual Modeling of Information Systems, Conceptual Modeling of Information Systems, 2007. doi:10.1007/978-3-540-39390-0.

[12] A. Bernasconi, et al., A comprehensive approach for the conceptual modeling of genomic data, in: Conceptual Modeling, Lecture Notes in Computer Science, Springer International Publishing, 2022, pp. 194–208.

[13] A. García, et al., The challenge of managing the evolution of genomics data over

time: a conceptual model-based approach, BMC Bioinformatics 23 (2022). doi:10.1186/s12859-022-04944-z.

[14] A. García, et al., 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.

[15] M. Costa, A. S., A. Palacio, A. Bernasconi, O. Pastor, A Reference Meta-model to Understand DNA Variant Interpretation Guidelines, 2023, pp. 375–393. doi:10.1007/978-3-031-47262-6_20.

[16] About the Unified Modeling Language Specification Version 2.5.1, https://www.omg.org/spec/UML/2.5.1/About-UML, 2017. (Accessed on 03/06/2024).