From Complexity to Clarity: Reproducible and Scalable Phenotype Development and application of LLM in a support role.

Asieh Golozar[1, 2], Albert Prats Uribe[3], Tom Seinen[4], Dani Prieto-Alhambra[3, 4], Peter Rijnbeek[4], Christian Reich[2, 4]

[1]Odysseus Data Services, Cambridge, MA, USA [2]OHDSI Center at the Roux Institute, Northeastern University, Boston, MA, USA [3] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK [4] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Introduction

Phenotyping aims at reliable and accurate identification of individuals by their observable traits from disparate observational data. These traits are conditions, observations, measurements and lab tests, diagnostic and therapeutic procedures, drug treatments, device applications, and encounter information with the healthcare system (visits). The accuracy of these phenotypic traits vary greatly. Conditions are particularly vulnerable as their recording is the result of a complex diagnostic process and subject to justification rules for reimbursement of healthcare provider's interventions. That can lead to overreporting (low specificity), underreporting (low sensitivity) and inaccurate timing of these records.

Therefore, to achieve accurate phenotypes, researchers cannot always rely merely on condition records (diagnostic codes). Instead, they combine them with additional data to boost the performance of the phenotype definition. For example, they ask for lab test results or therapeutic interventions specific to the condition. This turns phenotypes into complex algorithms consisting of Boolean and temporal logic.

Designing these phenotypes has fundamental challenges: They require in-depth understanding of the disease, its presentation, diagnosis, management and prognosis, and familiarity with medical practice across geographies and settings as well as reporting patterns to payers. This makes this process overly complex, time consuming and irreproducible. The performance of the definition and the contribution of each criterion are difficult to assess, since ground truth is only obtainable through lengthy validation from the charts.

Here, we introduce a structured and comprehensive process that guides researchers through the design and through the design process, addressing the limitations and arbitrariness of defining phenotype algorithms. It sets up the requirements for the phenotype and offers a systematic approach to the construction of criteria and their logical relationships. We believe this will make phenotypes development more transparent, parsimonious, efficient and can reduce the level of subjectivity throughout the process.

Methods

We developed a three-step process for developing what we call a specified phenotype.

Firstly, the requirements the specified phenotype are inferred from a few key dimensions. The dimensions fall into two broad categories, the disease presentation and the use case context. These define four requirements for optimization of the performance characteristics: of the sensitivity, specificity, index date and cohort end date (Table 1).

Table 1. The optimization requirement associated with each phenotype dimension.

Dimension	Disease dimension	Sensitivity	Specificity	Index date	End date
Disease presentation					
Record capture	Under-coded	optimize			
	Over-coded		optimize		
Disease pattern	Non-recurring				ignore
	Recurring			optimize	optimize
Use case context					
Severity	Severe		optimize		
	Grade/Stage		optimize		
Flavor	Incident			optimize	ignore
	Prevalent			ignore	optimize
Intended use	Exclusion	ignore	ignore	ignore	ignore
	Indication			optimize	
	Target			optimize	
	Outcome		optimize	optimize	
	Baseline characteristic		ignore		optimize
	Follow-up characteristic		ignore	optimize	

Secondly, a wire frame containing the index criterion, inclusion/exclusion criteria, entry and exit timing is created, all based on the previously determined requirements for the specified phenotype. For each such requirement, this process leads through a checklist of questions about the nature of the disease and its context, such as differential diagnosis, history of the disease, sequalae or complications. These questions guide driving the logic. If no optimization requirement is determined in step 1, the resulting phenotype consists just of an index criterion and an open cohort end. Otherwise, additional criteria and timing logic will be applied.

Finally, the details of the criteria and conditions are filled in, particularly the conceptsets.

To make this process feasible for researchers without deep knowledge of medical and administrative practice, we incorporated the advanced language models GPT-3.5 providing the medical knowledge. It will assist in step two and three using standardized prompts, answering the questions of the checklist. Additionally, we intend to utilize GPT3.5 to navigate through the hierarchies of the OHDSI Standardized Vocabularies and create conceptsets efficiently. We also plan to report on the performance and utility of GPT3.5 in phenotype development.

Results and Discussion

By adopting this structured approach, we aim to address the subjectivity and complexity of phenotype development while improving transparency, reproducibility, and efficiency. It will be particularly useful to the analyst without a medical background. The incorporation of advanced language models enhances the process by automating certain aspects, reducing subjectivity, and facilitating the creation of concept sets.

Conclusion:

The proposed standard approach for computable phenotype development provides a systematic and transparent framework for overcoming the challenges associated with subjective and labor-intensive methods.