GUEST EDITOR COMMENTARY

Why does human phenomics matter today?

Vasa Curcin

King's College London, London, UK

Correspondence
Vasa Curcin, King's College London, London, UK.
Email: vasa.curcin@kcl.ac.uk

Abstract
Human phenomics responds to an urgent need in the medical research community; namely, reproducibility.

KEYWORDS
computational phenotyping, electronic health record, human phenomics, patient phenotypes, provenance, reproducibility

Human phenomics is the science and practice of defining observable medical phenomena to advance research and personalised care. The concept of phenotype originated as a complement to the genotype, defined as a complete set of an individual's inheritable characteristics. Rather than describing someone’s genetic information, a phenotype captures all the observable properties that result from the interaction of their genetic make-up and environmental factors, for example, demographic information and medical events. With the emergence of large-scale Electronic Health Record (EHR) data repositories, the term phenotype has evolved to denote groups of patients sharing the same phenotypic traits, such as a set of patients suffering from some disease. In addition to disease phenotypes, it is also sometimes necessary to define a more complex cohort, including comorbidities, polypharmacy, and demographic data—we refer to these as patient phenotypes. Defining these phenotypes, and validating them to ensure their accuracy and generalizability, is the process known as phenotyping, with EHR-based phenotyping relying primarily on data in the patient's health record, and computational phenotyping using unsupervised machine learning techniques to discover novel phenotypes and investigate their properties.

Human phenomics responds to an urgent need in the medical research community, and that is the one of reproducibility, ensuring that published research is sufficiently detailed that its rigour and appropriateness can be reliably asserted and assessed. To achieve this trust in research outputs, the data from which they are derived needs to be of high quality, and also provenance and methods associated with its extraction and transformation processes must be fully documented.

The core challenge presented in this Special Issue is how to ensure that populations extracted from different datasets are clinically equivalent, and this is part of a broader conversation that is ongoing in a number of communities. While phenotypes are the cornerstone of any observational study that utilises EHR-derived datasets, and health service applications where well-defined patient cohorts are used as numerators or denominators of quality metrics, any field relying on a strict cohort characterisation requires precise, portable phenotype definitions. In clinical trials, detailed eligibility criteria are needed for the successful recruitment of trial sites, by querying site EHR repositories to determine whether there are significant numbers of a certain phenotype present in the patient population. Decision support systems embedded in EHR systems are triggered when a set of conditions is satisfied, that is, when a patient conforms to the phenotypical definition. More broadly, the application of any predictive model assumes that the model's preconditions can be mapped onto the target dataset, a task that is impossible without reliable equivalences between the model source data and the novel data to which it is applied.

Early approaches to defining phenotypes consisted of lists of SNOMED or ICD codes for particular diseases, which cannot adequately capture nuances of multi-morbidities and patient demographics. The portability of data element definitions is now increasingly achieved through standards such as Observational Medicine Outcome Partnership Common Data Model (OMOP CDM)1 which map onto multiple coding standards, but in themselves are still not sufficient to describe complex phenotypes which require additional logic, for example, temporal sequencing of events in patient histories, or use of external tooling, such as natural language processing. Thus, a logical layer is needed to orchestrate the data elements. OMOP's Atlas tool uses webform templates for the design of OMOP-based phenotypes based on multiple clinical events with temporal constraints between them. Other phenotyping tools, that are agnostic in their choice of data model, use generic programming techniques such as scripting languages or scientific workflows.

© 2020 The Author. Learning Health Systems published by Wiley Periodicals LLC on behalf of the University of Michigan.
The portability of computable phenotype definitions, however well-defined, is of little use without effective dissemination strategies that allow researchers to discover and access curated and validated phenotypes. Initial efforts in building standardised phenotype repositories, such as the Phenotype Knowledge Base (PheKB), UK’s CALIBER,3 Million Veterans Program (MVP),4 and All of Us consortium5 have attracted thousands of users within their research programmes. In the United Kingdom, the Health Data Research UK network is building the National Phenomics Resource (https://www.hdruk.org/projects/national-phenomics-resource/) based on technologies developed in CALIBER, Sail Databank6 and Phenoflow architecture.7

1 | THE STATE OF RESEARCH IN PHENOMICS: WHAT THIS SPECIAL ISSUE TELLS US

The separation of design and execution is one of the main research efforts in phenotype research. Brandt et al8 investigate the suitability of Clinical Quality Language (CQL) for computable phenotype definitions of arbitrary complexity and validate the approach through cross-database and cross-platform executions of the phenotype.

Patient-centred outcomes (PCOs) are recognised as a challenge for phenotyping due to the lack of relevant structured information that can serve as the basis for phenotyping algorithms. Hernandez-Boussard et al9 suggest natural language processing (NLP) as a mechanism for integrating PCOs into phenotype definitions in a structured manner. Severity-based phenotypes introduce additional complexity, which the authors address by combining NLP with rule-based models.

Determining whether a phenotyping algorithm can be applied to a dataset is not only a methodological task, but also a data quality issue and mechanisms are needed to test data sets for fitness of purpose with respect to a particular algorithm, particularly when portability occurs between health settings. Wiegersma et al10 look into the methods for such testing on the example of Sjögren’s syndrome code list-based phenotype algorithm developed from the Dutch national primary care database and applied to a hospital insurance claim database.

Computational phenotyping, where machine learning algorithms are employed to learn novel patient groupings, is a growing area and a subject of three papers in this Special Issue. Z. Xu et al11 demonstrate an application of a computational technique for learning different depression phenotypes and investigating them to assert their properties in terms of observed characteristics. A similar method is used in J. Xu et al12 to identify subphenotypes of Alzheimer’s Disease, a notoriously difficult condition to extract from EHRs, with the aim of discovering clusters of patients with specific characteristics that could be targeted for personalised treatments. Finally, development of a computable phenotype for Crohn’s Disease (CD) is the subject of Dempsey et al,13 whereby OMOP-mapped code lists are used to determine a CD phenotype from PEDSnet network data sources, and validated against a national registry of CD patients.

2 | FUTURE DIRECTIONS FOR RESEARCH ON PHENOMICS

The papers in this Special Issue collectively identify a number of future directions in human phenomics research that need to be explored to support the application and re-use of phenotypes in various settings. Human phenomics is rapidly developing but is still at an early stage where methods and repositories are emerging to meet the needs of a range of medical research domains. As tooling gradually matures beyond the realm of enthusiastic software to become usable for a broad spectrum of researchers and implementers, the focus needs to move to techniques for moving phenotype definitions between data sources, but also health settings. Many questions still need to be answered: Can we design phenotypes which can apply in a primary care physicians’ practice, and in an intensive care unit? Can we layer these algorithms in such ways that some degree of similarity is retained?

While at present much of the phenomics work is focused on stand-alone phenotypes, relationships between related phenotypes are an important area worth exploring in the context of phenotype warehouses, especially as many chronic conditions co-occur. Sub-phenotypes may inherit the properties of their parents, but no repository currently models this. Similarity metrics between phenotype definitions will assist in scalable searches across different repositories, whereby a partial match may indicate a usable cohort definition to investigate.

Validation at scale is another area that urgently requires research attention. Historically, individual disease registries have been used as gold standards for phenotypes related to those diseases, but this is not scalable or feasible for patient cohorts focusing on multimorbidities and complex demographic criteria. Clinical notes reviews, where phenotype patient matches are manually reviewed, are not sustainable for large LHS infrastructures. While manual text extraction of phenotypes can be effective in smaller scenarios, it is heavily dependent on the human expert and the sample being analysed, and not well-suited to cross-site studies with differences in clinical and operational procedures between sites. Novel hybrid approaches that encompass structured data, free text and ancillary sources for both structured and unstructured data are badly needed to validate phenotypes efficiently and reliably at scale.

3 | CONCLUSION

The articles in this Special Issue represent a variety of robust efforts that collectively describe the importance of human phenotyping and state of the art in the field. We highlight several needs and opportunities for future work at the intersection of learning health systems and human phenomics, including a need for formal computable definitions, methods for validation and evaluation of published phenotypes, and routes to addressing local variability in operational practice. Perhaps most importantly, we stress the need for using these advances to
support sharing of research methods leading to increased trust and stronger impact.

Although we are excited to present the collection of articles in this Special Issue, much work remains to address the knowledge gaps in the field. We invite authors to continue submitting articles focusing on phenomics and computational phenotyping to Learning Health Systems.

ORCID

Vasa Curcin https://orcid.org/0000-0002-8308-2886

REFERENCES

1. The Book of OHDSI: Observational Health Data Sciences and Informatics. https://www.hdruk.org/projects/national-phenomics-resource/ (n.d.).
2. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016;23(6):1046-1052. https://doi.org/10.1093/jamia/ocv202.
3. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UKphenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019;26(12):1545-1559. https://doi.org/10.1093/jamia/ocz105.
4. Knight KE, Honerlaw J, Danciu I, et al. Standardized architecture for a mega-biobank phenomic library: the million veteran program (MVP). AMIA Jt Summits Transl Sci Proc. 2020;326-334.
5. The All of Us Research Program Investigators. The “all of us” research program. N Engl J Med. 2019;381:668-676. https://doi.org/10.1056/NEJMsr1809937.
6. Jones K, Ford D, Jones C, et al. A case study of the secure anonymous information linkage (SAIL) gateway: a privacy-protecting remote access system for health-related research and evaluation. J Biomed Inform. 2014;50:196-204. https://doi.org/10.1016/j.jbi.2014.01.003.
7. Chapman M, Rasmussen LV, Pacheco JA, Curcin V. Phenoflow: portable workflow-based phenotype definitions. medRxiv. 2020. https://doi.org/10.1101/2020.07.01.20144196.
8. Brandt PS, Kiefer RC, Pacheco JA, et al. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. Learn Health Sys. 2020:e10233. https://doi.org/10.1002/lrh2.10233.
9. Bozkurt S, Paul R, Coquet J, et al. Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. Learn Health Sys. 2020:e10237. https://doi.org/10.1002/lrh2.10237.
10. Wiegensma S, Flinterman LE, Seghieri C, et al. Fitness for purpose of routinely recorded health data to identify patients with complex diseases: The case of Sjögren’s syndrome. Learn Health Sys. 2020; e10242. https://doi.org/10.1002/lrh2.10242.
11. Xu Z, Wang F, Adekkannattu P, et al. Subphenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. Learn Health Sys. 2020; e10241. https://doi.org/10.1002/lrh2.10241.
12. Xu J, Wang F, Xu Z, et al. Data-driven discovery of probable Alzheimer’s disease and related dementia subphenotypes using electronic health records. Learn Health Sys. 2020;e10246. https://doi.org/10.1002/lrh2.10246.
13. Khare R, Kappelman MD, Samson C, et al. Development and evaluation of an EHR-based computable phenotype for identification of pediatric Crohn’s disease patients in a National Pediatric Learning Health System. Learn Health Sys. 2020;e10243. https://doi.org/10.1002/lrh2.10243.

How to cite this article: Curcin V. Why does human phenomics matter today? Learn Health Sys. 2020;4:e10249. https://doi.org/10.1002/lrh2.10249