Computable phenotypes for cohort identification: core content for a new class of FAIR Digital Objects

Marisa L Conte‡, Allen J Flynn‡, Peter Boisvert‡, Zach Landis-Lewis‡, Rachel L Richesson‡, Charles P Friedman‡

‡ Department of Learning Health Sciences, Medical School, University of Michigan, Ann Arbor, United States of America

Corresponding author: Marisa L Conte (meese@umich.edu)

Abstract

Introduction

We present current work to develop and define a class of digital objects that facilitates patient cohort identification for clinical studies, such that these objects are Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al. 2016). Developing this class of FAIR Digital Objects (FDOs) builds on the work of several years to develop the Knowledge Grid (https://kgrid.org/), which facilitates the development, description and implementation of biomedical knowledge packaged in machine-readable and machine-executable formats (Flynn et al. 2018). Additionally, this work aligns with the goals of the Mobilizing Computable Biomedical Knowledge (MCBK) community (https://mobilizecbk.med.umich.edu/) (Mobilizing Computable Biomedical Knowledge 2018). In this abstract, we describe our work to develop a FDO carrying a computable phenotype.

Defining computable phenotypes

In biomedical informatics, 'phenotyping' describes a data-driven approach to identifying a group of individuals sharing observable characteristics of interest, generally related to a disease or condition, and a 'computable phenotype' (CP) is a machine-processable expression of a phenotypic pattern of these characteristics (Hripcsak and Albers 2018).

For the purposes of this work, we are interested in CPs derived from data contained in electronic health record (EHR) systems. This includes both structured data, e.g. codes for diseases, diagnoses, procedures, or laboratory tests, and unstructured data, e.g. free text including patient histories, clinical observations, discharge summaries, and reports. Thus, we define computable phenotype FDOs (CP-FDOs) as a class of FDO that packages an executable EHR-derived CP together with documentation needed to implement and use it effectively for creating cohorts of individuals with similar observable characteristics from EHR data sets.

Importance of portable and FAIR CPs

There is tremendous excitement for using real-world EHR data to discover important findings about human health and well-being. However, for discovery to happen, researchers need mechanisms like CPs to identify study cohorts for analysis. Beginning in the early 2010s, a growing literature explores various methods for the secondary use of EHR data for patient phenotyping to arrive at consistent study cohorts (Shivade et al. 2014, Banda et al. 2018). The heterogeneous nature of EHR data has inspired a wide variety of phenotyping methods, from those which rely solely on documented codes linked to terms in existing vocabularies to those which combine such codes with other concepts extracted from free text using natural language processing.

Our current focus is on packaging CPs inside FDOs for classifying patients as having or not having a phenotype of interest. This can be done within an individual health system, or at scale across a clinical data research network. Using CPs for cohort identification can reduce the time and expense of traditional data set building and clincal trial recruitment, and expand the potential scope of a study population(Boland et al. 2013).

Creating and validating CPs requires time, resources, and both clinical and technical expertise. One estimate is that it can take 6-10 months to develop and validate a CP (Shang et al. 2019). And, as there is no standard data model within EHRs in the United States, many CPs are designed for performance at a single site, rather than for portability, which is understood as the ability to implement a phenotype at a different site with similar performance (Shang et al. 2019). While portability is increasingly recognized as an important element of phenotyping, and there have been recent efforts to develop more portable CPs, many of these processes still require significant technical expertise at the implementation site to adapt the phenotype for use on local data.

There may also be significant advantages to making CPs FAIR. These include transparency in cohort selection, and better generalizability of results. FAIR CPs may also increase the potential for robust comparisons of data from related studies, leading to better evidence synthesis to improve delivery of care and ultimately human health.

Defining a new class of FDOs to hold and convey CPs

We believe that packaging validated CPs inside digital objects may alleviate many of the pressures mentioned above, and contributes to making both the processes and products of clinical research more FAIR. To this end, our current work focuses on packaging a validated CP inside a machine-processable FDO. The phenotype of interest identifies pediatric and adult patients with a rare disease (Oliverio et al. 2021), and has several features which make it ideal for transformation to an executable FDO. First, the phenotype utilizes standards to define the clinical characteristics of interest, and is based on a common data model; these features increase the potential for both interoperability and reuse. Additionally, because the phenotype has been validated across three sites, its portability has already been demonstrated. Finally, the full computable phenotype has been shared as a series of SQL queries, including scripts for patient identification, deriving

statistics, and validation, which have been annotated with instructions for implementation at other sites.

The goals of this work are:

1. To develop CPs as executable DOs, leveraging previous work to develop executable Knowledge Objects (KO) (Flynn et al. 2018)
2. To advance our understanding of how to define computable phenotypes as a class of FDO, including what is needed to meet the requirements of binding, abstraction, and encapsulation (Wittenburg et al. 2019)

Conclusion

Computable phenotypes, packaged as FDOs, may increase the potential both for the portability of a phenotype and the reusability of data resulting from its implementation. Providing CPs as executable FDOs may also reduce barriers to portability and local implementation. In this presentation, we describe our work to develop a FDO computable phenotype from an existing validated phenotype. Lessons learned from this process will increase our understanding of both the technical requirements, and how to address necessary components of abstraction, binding, and encapsulation so that these can function as FAIR Digital Objects.

Keywords

computable biomedical knowledge, portability, reuse

Presenting author

Marisa L Conte

Presented at

First International Conference on FAIR Digital Objects, presentation

Conflicts of interest

References

- Banda J, Seneviratne M, Hernandez-Boussard T, Shah N (2018) Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. Annual Review of Biomedical Data Science 1 (1): 53-68. https://doi.org/10.1146/annurev-biodatasci-080917-013315

- Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C (2013) Defining a comprehensive verotype using electronic health records for personalized medicine. Journal of the American Medical Informatics Association 20 https://doi.org/10.1136/amiajnl-2013-001932
- Flynn A, Boisvert P, Gittlen N, Gross C, Iott B, Lagoze C, Meng G, Friedman C (2018) Architecture and Initial Development of a Knowledge-as-a-Service Activator for Computable Knowledge Objects for Health. Studies in Health Technology and Informatics 247: 401-405.
- Hripcsak G, Albers DJ (2018) High-fidelity phenotyping: richness and freedom from bias. Journal of the American Medical Informatics Association 25 (3): 289-294. https://doi.org/10.1093/jamia/ocx110
- Mobilizing Computable Biomedical Knowledge (2018) MCBK Manifesto. URL: https://mobilizecbk.med.umich.edu/about/manifesto
- Oliverio A, Marchel D, Troost J, Ayoub I, Almaani S, Greco J, Tran C, Denburg M, Matheny M, Dorn C, Massengill S, Desmond H, Gipson D, Mariani L (2021) Validating a Computable Phenotype for Nephrotic Syndrome in Children and Adults Using PCORnet Data. Kidney360 2 (12): 1979-1986. https://doi.org/10.34067/KID.0002892021
- Shang N, Liu C, Rasmussen L, Ta C, Caroll R, Benoit B, Lingren T, Dikilitas O, Mentch F, Carrell D, Wei W, Luo Y, Gainer V, Kullo I, Pacheco J, Hakonarson H, Walunas T, Denny J, Wiley K, Murphy S, Hripcsak G, Weng C (2019) Making work visible for electronic phenotype implementation: lessons learned from the eMERGE network. Journal of biomedical informatics 99 https://doi.org/10.1016/j.jbi.2019.103293
- Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM (2014) A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association 21 (2): 221-230. https://doi.org/10.1136/amiajnl-2013-001935
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Wittenburg P, Strawn G, Mons B, Bonino L, Schultes E (2019) Digital Objects as Drivers towards Convergence in Data Infrastructures. web publication https://doi.org/10.23728/B2SHARE.B605D85809CA45679B110719B6C6CB11