Analysis of Biomarker Data Towards Development of a Molecular Biomarker Ontology

Daniel Lyman[1], Darren Natale[2], Lynn Schriml[3], Kriston Anton[4], Daniel C. Crichton[5], Raja Mazumder[1]

[1] *The George Washington University, 2121 I St NW, Washington, DC, USA*
[2] *Georgetown University Medical Center, 37th and O Street, N.W., Washington, DC, USA*
[3] *University of Maryland School of Medicine, 655 W Baltimore St S, Baltimore, MD, USA*
[4] *University of North Carolina, Chapel Hill, NC, USA*
[5] *NASA Jet Propulsion Laboratory, 4800 Oak Grove Dr, Pasadena, CA, USA*

Abstract
Molecular biomarkers comprise fundamental elements of biomedical inquiry. No ontology has been developed to organize this knowledge across disciplines and to link diseases, processes, or technologies, which would be a substantial asset for research and healthcare. A sustained effort is under way to construct such an ontology for greater precision in representation. Observed overlaps of biomarkers for COVID-19, diabetes, and cancers underscore the potential of novel ontology-based explorations. A biomarker ontology can harmonize data across varied diseases and technologies, tie together NIH programs, guide future data collection, support machine learning, and foster research in ethical use of biomarkers.

Keywords
Ontology development, molecular biomarkers, data integration, data linkage, disease

1. Introduction

Acquisition and use of knowledge from biomedical research reduce illness and enhance human health. Investigations of living systems routinely assess data on objects used as indicators (i.e., biomarkers) of biological processes, at the molecular, cellular, or physiologic level. The FDA-NIH Biomarker Working Group (BEST Resource [1]) defines biomarkers as "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or biological responses to an exposure or intervention, including therapeutic interventions". As indicators of biological processes, biomarkers comprise a growing focus of biomedical research: as crucial factors in biological inquiries, essential elements of precision medicine, critical components of pipeline screens or clinical trials, and vital ingredients of investment decision making in the development of therapeutics.

Evidence indicates that impaired cellular components contribute to the onset or progression of many disorders; hundreds of genes and proteins, for example, are implicated in oncogenic pathway deregulations. Advances in omics technologies have driven the use of molecular biomarkers as principal instruments of current inquiry and improvements in medical treatment. Molecular biomarkers gauge and illuminate alterations of specific genomic, proteomic, glycomic, lipidomic, or metabolomic components that underlie key functions. Biomarkers, therefore, comprise a fundamental nexus of biomedical inquiry and great interest exists across medical disciplines in further biomarker discoveries.

Accordingly, numerous databases collect biomarker content; e.g., OncoMX, EDRN (Early Detection Research Network), the Alzbiomarker DB, MarkerDB, and others. Additional databases collect closely related content. However, these resources are not harmonized with a common standard.

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy
EMAIL: danlyman@email.gwu.edu; dan5@georgetown.edu; lschriml@som.umaryland.edu; kristen_anton@med.unc.edu; daniel.j.crichton@jpl.nasa.gov; mazumder@gwu.edu
ORCID: 0000-0001-9333-1455; 0000-0001-5809-9523; 0000-0001-8910-9851; 0000-0002-5487-7719; 0000-0001-8823-9945

Despite their central investigative importance (Figure 1), no harmonized organization (vocabulary or ontology) of biomarker knowledge has been developed as a cross-cutting infrastructure or unifying instrument across diseases, processes, components, or technologies. As a result, significant challenges hinder generation and translation of integrated biomarker observations into beneficial clinical applications. A logic-based data structure that facilitates pan-biomarker investigation would be a significant asset. Formal models have enhanced analysis of large datasets; identified driver genes and determinants; supported patient staging; identified treatment options; and predicted responses to therapy and survival. However, a recent search of resources and literature identified only narrowly-scoped biomarker ontologies; e.g., Imaging Biomarker Ontology [2], Food-Biomarker Ontology [3], and Coronavirus Infectious Disease Ontology [4]. Other identified ontologies are inactive or not publicly available and the Ontology for General Medical Science [5], describing clinical encounters, has been extended with definition and classification of disjoint biomarker types (material, quality, process), but does not include processes of biomarker measurements.

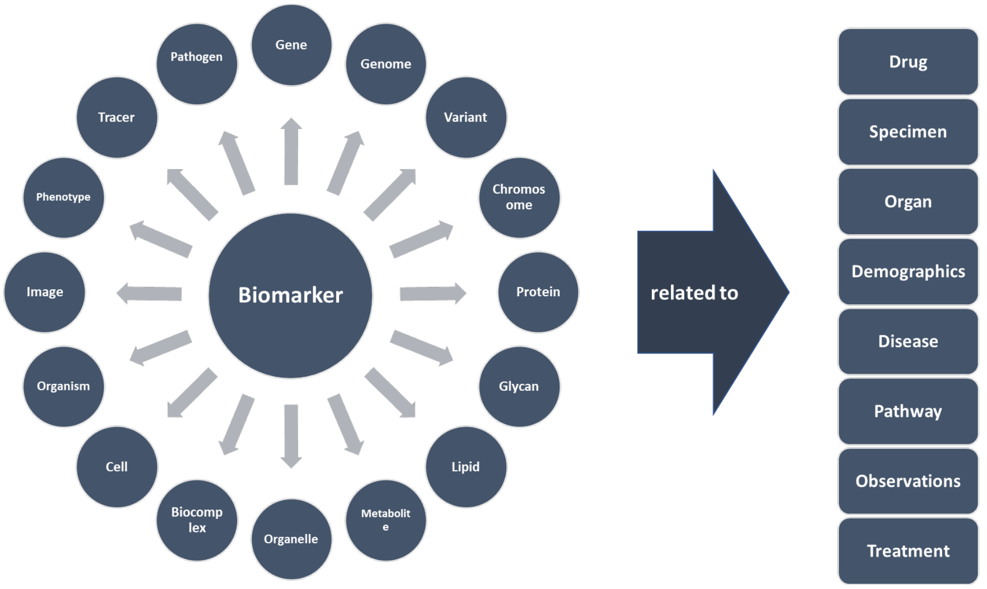

Figure 1: Molecular biomarkers may comprise assorted interrelated biological objects (circle) that are also *related to* other biomedical objects, properties, or processes (column) in particular ways that can be specifically described and represented in a biomarker ontology.

2. Results and Discussion

Through work on biomarker discovery, data integration, and ontology development [6-9], we see an urgent need to harmonize key biomarker knowledge, organized on OBO Foundry principles [10] and linked with related models, for cross-disciplinary investigations and exploration of novel hypotheses. A sustained effort is, therefore, under way to evaluate biomarker knowledge towards identifying and defining domain scope, classes, attributes, and relations. The goal is to harmonize terminology, structured in a machine-readable framework, and axiomatically connected to related elements: e.g., gene, protein, disease, phenotype, cell, anatomy, variants. The general workflow of the project is shown in Figure 2.

Following literature reviews, searches of public resources, and discussions with associated parties, the scope of the initial model has focused on 1) molecular disease biomarkers; 2) primary use cases; 3) examination of source materials; 4) ontological evaluation of biomarker objects; and 5) evaluation of associations across related elements. Examination of data models and resource contents has provided an outline of the subject landscape, domain elements, and connections with related data. Detailed inspection of source materials has refined our view of biomarker and related knowledge: metadata, biomarker instances and measures, definitions, and labels, as well as classification and relations. Collection of essential biomarker data types, terms, classes, and relations from databases, resources, and use cases provides elements to design a foundational representation and informs the import of data from use case resources. Use cases provide practical substance to devise representation and organization of data in an initial model that addresses known biomarkers (single and panels) and data types, is able to evolve to a material ontology, and can accommodate future data, data types, and concepts.

Rapid publication of studies in 2020, proposing new COVID-19 biomarkers for therapy, evaluation, risk assessment. etc., served as one case that highlighted the need for biomarker harmonization and integration. It further provided an opportunity to test a data model to capture, describe, and accommodate key data as new biomarkers were identified. Crowdsourced curation of COVID-19 studies extracted 188 biomarkers (gathered to date) that facilitated configuration of a model comprised of core attributes: Assessed Biomarker Entity, Biomarker Measurement, Cross Reference to a standard resource (e.g., UniProt accession [11]), BEST Biomarker Type [1], Specimen Type (with Uberon ID), Disease Name (with DOID), LOINC Code, Literature Evidence (declarative source text), and Notes (further information). Each annotated biomarker entry is assigned a unique alphanumeric ID and labels for Assessed Biomarker Entity and the Biomarker Measurement are standardized.

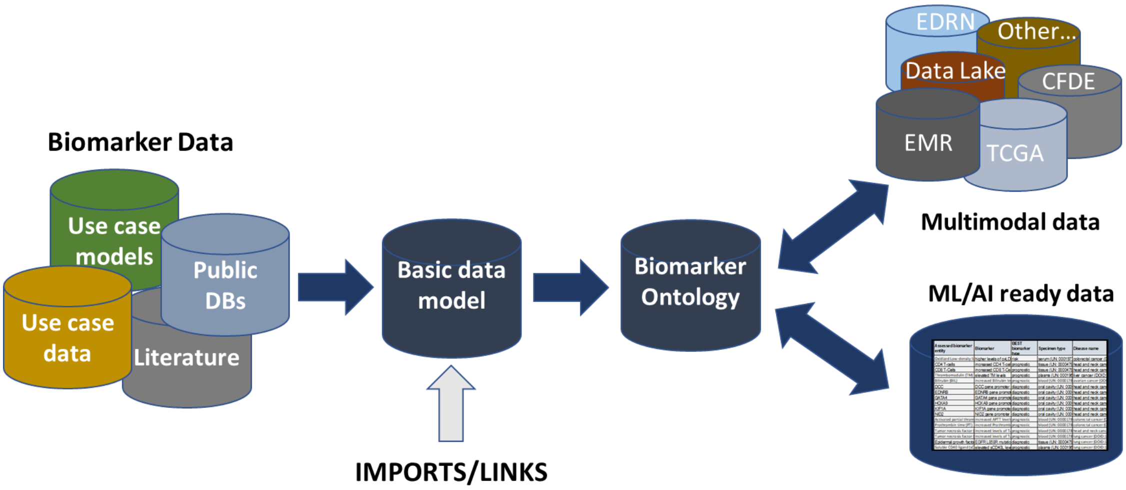

Figure 2: Biomarker data is collected from assorted resources, organized, and stored in a basic, extensible data model. Following ontological analysis and import of related knowledge, stored biomarker data is reformatted into a machine-readable ontology for efficient and flexible data query, sharing, and analysis across biomarker and related domain knowledge for generation of novel hypotheses, exploration, and discovery.

Analysis of the biomarker data found increased or decreased activity of specific proteins, metabolites, lipids, glycoconjugates, carbohydrates, electrolytes, and circulating blood cells [12]. Examination of biomarkers indicated COVID-19 responses to viral challenge, immune activation and regulation (innate and adaptive), and effects in cell growth, coagulation, vascular remodeling, homeostasis, cell adhesion, and metabolism. Thirty of the biomarkers are also biomarkers of diabetes and showed similar increased or decreased biologic activities in important host functions. Eighty-eight biomarkers are also biomarkers of twenty-three different cancer types and involve cell growth and proliferation, homeostasis, metabolic components, cell adhesion, coagulation, and immune activation and regulation. These overlaps underscore the need to harmonize biomarker knowledge across diseases, resources, and technologies and suggest potential ontology-based discovery of significant biomarkers through exploration of key pathways, cross-associations among diseases, and novel avenues of investigation.

We are also curating biomarkers for 15 broad cancer types (373 biomarkers gathered to date), which show altered activity of proteins, genes, metabolites, lipids, carbohydrates, and cells. Several entities are biomarkers for more than one cancer and seven biomarkers of diabetes are also biomarkers of cancers, involving primarily immune activation and regulation. The cancer biomarker data is supplemented with mappings of single (1600) and panel gene/protein biomarker data from EDRN and FDA. Curation of N-linked glycans in hepatocellular carcinoma facilitates further cross-cutting analyses, establishing a unique environment for glycan panel data linked with genomic and proteomic data, which may differentiate N-glycosylation profiles in cancer cohorts. All data collected to support evolution of a molecular biomarker ontology are freely available for download and analysis at https://data.oncomx.org/.

Analysis of biomarker data collected from diverse sources supports acquisition and detection of shared ontological attributes that express properties of components in the knowledge space; e.g., harmonization of terms, expressions, and semantics; identification, definition, and organization of classes and relations. Ontological analysis and representation of biomarker data began with, and rests upon, the BEST Resource definition of biomarkers [1], which includes features described in other definitions. Rigorous examination of the definition indicates (unstated, though implied) that each instance of a measured indicator object is compared to an established standard measure of that object, to assess an instance of a biological process (or response). The definition further implies that measurement of the indicator object (the observed result compared to a standard), rather than the indicator object per se, constitutes the biomarker. To illustrate the ground semantics, an increased level of a protein (rather than the protein) is a biomarker; a sequence variant of a gene (rather than the wild type gene) is a biomarker; altered structure of a glycan (rather than the typical glycan) is a biomarker; and so on.

However, conventional designations about biomarkers in databases, publications, and data models customarily name the standard referent substance (e.g., 'IL-6' or 'BRCA1') as an easily recognized label (understood as proxy) for the measured object, instead of precisely naming the actual measured object. Ontological implementation of this common name conflation (biomarker and referent), however, can produce logical errors. Our approach explicitly expresses in the model the precise intended semantics in a biomarker name/label (e.g., 'increased IL-6 expression' or 'BRCA1 variant xyz'). Representation of knowledge in the emerging ontology, assembled from diverse resources, therefore, includes a distinction between "Assessed Biomarker Entity" and "Biomarker". An "Assessed Biomarker Entity" (e.g., 'IL-6') in the ontology designates the standard referent object, while named "Biomarker" objects (e.g., 'increased IL-6 expression') designate measured indicator objects.

Analysis of biomarker content from diverse sources has collected assorted examples of biomarker types, identified diverse relations to facilitate semantic reasoning, created annotations, and distinguished links to entities in related models. Biomarker terms will be hierarchically organized into classes, subclasses, and instances accordingly. Ontological analysis indicates that biomarkers may be classified along several possible axes: for example, 1) intended clinical use of a biomarker with respect to some disease/medical condition or medical product/environmental agent; 2) specimen source of a biomarker; 3) type of assessed entity of a biomarker. We have chosen assessed entity as an asserted classification axis, since each biomarker can appear in only one sub-branch of this axis and we have begun to examine custom relations with ranges set to ensure proper reasoning; such as indicated_by_increased_level_of for some biomarkers X logically defined as elevated X level (and converse) or with indicated_by_increased_level_of for some biomarkers defined as computed ratios, for example. We have also examined the use of relations to classify biomarkers by reasoning in non-asserted axes and successfully tested that reasoning performs correctly (including also a 'fake' example that catches errors in logic). Relations will be further reviewed with respect to the Relations Ontology [13]. Analysis of molecular biomarker source materials, data models, and resource contents has also prompted closer ontological investigation of Biomarker relations to Phenotypes and Risk Factors.

Evolution of the model will be informed by data collected from biomarker databases, use case data models, and peer-reviewed publications, including novel data types. Additional data types will also be obtained from biomarkers undergoing clinical trials or those approved by the FDA. Additional use cases are of interest. Related data on genes, proteins, phenotypes, cells, and anatomy will be identified in reference ontologies, comprising critical resources for cross-disciplinary interoperability and integration and providing key components for harmonization, alignment, and relations. To enhance the core model, we will import, integrate, and connect (via axioms) data of selected reference ontologies for domain coverage, data sharing, and exploration of knowledge in relevant resources. Annotations with imported data will reveal connections between biomarkers associated with specific diseases and phenotypes; with underlying protein and gene actors; and with affected cells, tissues, and organs.

3. Conclusion

Ontological modeling will provide a standardized terminology, structured representation of molecular biomarker data and knowledge, consistent rich annotations of biomarker objects in machine-readable language, and logically inferable knowledge for data science approaches to discovery. Explicit assertions of common properties will link biomarker elements with knowledge in related models, datasets, and nodes (e.g., protein, disease, phenotype, anatomy, and more), facilitating data integration and interoperability across critical independent data resources through the lens of biomarker associations. Structured representation of molecular biomarker knowledge will promote efficient, FAIR [14], and flexible data query, acquisition, sharing, and analysis. Unification of data across the biomarker domain and with related knowledge enables generation of novel hypotheses, exploration, and discovery.

4. Acknowledgements

Work was supported in part by National Cancer Institute (NCI) (U01CA0215010) to RM and DC.

5. References

[1] FDA-NIH Biomarker Working Group, BEST (Biomarkers, EndpointS, and other Tools) Resource, 2016. URL: https://www.ncbi.nlm.nih.gov/books/NBK326791.

[2] E. Amdouni, B. Gibaud, Imaging Biomarker Ontology (IBO): A Biomedical Ontology to Annotate and Share Imaging Biomarker Data, J. Data Semantics. 7 (2018) 223. doi:10.1007/s13740-018-0093-3.

[3] P. Castellano-Escuder, R. González-Domínguez, D. S. Wishart, C. Andrés-Lacueva, A. Sánchez-Pla, FOBI: an ontology to represent food intake data and associate it with metabolomic data, Database, 2020 (2020) baaa033. doi:10.1093/databa/baaa033.

[4] Y. He, H. Yu, E. Ong, et al., CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis, Sci. Data. 7 (2020) 181. doi:10.1038/s41597-020-0523-6.

[5] W. Ceusters, B. Smith, Biomarkers in the ontology for general medical science, Stud. Health Technol. Inform. 210 (2015) 155-9. doi:10.3233/978-1-61499-512-8-155.

[6] D. J. Crichton, A. Altinok, C. I. Amos, et al., Cancer Biomarkers and Big Data: A Planetary Science Approach, Cancer Cell. 38 (2020) 757-760. doi:10.1016/j.ccell.2020.09.006.

[7] H. M. Dingerdissen, F. Bastian, K. Vijay-Shanker, et al., OncoMX: A Knowledgebase for Exploring Cancer Biomarkers in the Context of Related Cancer and Healthy Data, JCO. Clin. Cancer Inform. 4 (2020) 210-220. doi:10.1200/CCI.19.00117.

[8] D. A. Natale, C. N. Arighi, J. A. Blake, et al., Protein Ontology (PRO): enhancing and scaling up the representation of protein entities, Nucleic Acids Res. 45 (2017) D339-D346. doi:10.1093/nar/gkw1075.

[9] L. M. Schriml, E. Mitraka, J. Munro, et al., Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic Acids Res. 47 (2019) D955-D962. doi:10.1093/nar/gky1032.

[10] B. Smith, M. Ashburner, C. Rosse, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (2007) 1251-5. doi:10.1038/nbt1346.

[11] UniProt consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506-D515. doi:10.1093/nar/gky1049.

[12] N. Gogate, D. Lyman, A. Bell, et al., COVID-19 biomarkers and their overlap with comorbidities in a disease biomarker data model, Brief Bioinform. (2021) bbab191. doi:10.1093/bib/bbab191.

[13] B. Smith, W. Ceusters, B. Klagges, et al., Relations in biomedical ontologies, Genome Biol. 6 (2005) R46. doi:10.1186/gb-2005-6-5-r46.

[14] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data. 3 (2016) 160018. doi:10.1038/sdata.2016.18.