Biomarker Partnership Data Model

CFDE-Draft-# (RFC Admin Only):	CFDE-DRAFT-12		
CFDE-Draft-Title:	Biomarker Data Model Implementation in the Partnership		
CFDE-Draft-type: (Right click the box and select checkmark ✔ CTRL - Z to undo)	Standard Implementation ✓	Design Principle ❏	Policy ❏
Point of Contact Name:	Daniall Masood		
Point of Contact Email Address:	daniallmasood@gwu.edu		
End of Date Comments *(MM/DD/YY)*:	11/20/2023		
Submitting Team/DCC Name:	GlyGen		
CFDE-Draft-Status:	Pending ☑	Active and Open for Comments □	
URL Link to the document:	https://docs.google.com/document/d/1vsoq4Z2yRA2Ee-CLod7J56dFwH9N3OLdwI-PPNYkY/edit		
License (Optional):			

Purpose

The purpose of the data model is to enforce a standardized structure and format for incoming data before it is incorporated. This will ensure that the incoming data is valid for the downstream processes of assigning unique ID's and processing for the ontology. The data model being proposed is based on the ontology structure that Darren Natale is developing. Additional fields can be added to the data model but current fields cannot be removed as it is the minimum data required for the ontology to work.

Biomarker Definition

The biomarker definition according to the FDA-NIH Biomarker Working Group (FNBWG) is “characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions.”

This means that when looking at a biomarker it should not be an entity by itself but an entity that exhibits a change that moves away from normal function and could indicate disease or a response to an exposure or intervention. The biomarker data model is centered around this definition.

Abbreviations

FNBWG: FDA-NIH Biomarker Working Group (https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/fda-biomarkers-working-group)
UPKB: Uniprot Knowledgebase (https://www.uniprot.org/)
DOID: Disease Ontology ID (https://www.disease-ontology.org/)
LOINC: Logical Observation Identifiers Names and Codes (https://loinc.org/)
BEST: Biomarkers, EndpointS, and other Tools (https://www.ncbi.nlm.nih.gov/books/NBK326791/)
HPO: Human Phenotype Ontology (https://hpo.jax.org/app/)
PMID: PubMed Identifier (https://pubmed.ncbi.nlm.nih.gov/)
dbSNP: Database for Single Nucleotide Polymorphisms (https://www.ncbi.nlm.nih.gov/snp/)

Data Model v1.0

Textual descriptions

‘biomarker’ field takes in information that follows the definition and the next fields provide contextual details that are directly related to the biomarker that is placed in this field. So the information in this field will be the change that is occurring in an entity to make it a biomarker.
Ex: “presence of rs1800562 mutation”

‘assessed_biomarker_entity’ provides the biomarker entity name where the change occurs along with a biological named entity term (gene name, protein name, etc). It is important to note that the main difference between this and ‘biomarker’ is the indication of the change that is occurring in the entity.
Ex: “Interleukin-6 (IL6)”

‘assessed_biomarker_entity_ID’ is simply the accession or identifier that is commonly used for the biomarker entity and is an outside source where the user can be pointed towards to learn more about the entity.
Ex: “UPKB:P05231”

‘assessed_entity_type’ is the entity type of the biomarker provided (e.g. gene, protein, glycan, cell).
Ex: “glycan”

‘best_biomarker_role’ provides the BEST category the biomarker belongs to as according to the definitions of BEST categories (BEST biomarker glossary) that are defined by the FDA-NIH Biomarker working group. These fields are required to be populated into the database as they contribute essential information towards the biomarker provided as well as the biomarker definition.
Ex: “diagnostic_biomarker”

‘Condition’ and ‘condition_ID’ are conditionally required fields in this data model. If the 'condition' field is provided, 'condition_ID' is required, and vice versa. Not all biomarker types are associated with a condition. Response, safety, and predictive biomarkers are associated with the response or reaction that a patient has to an environmental agent or medical intervention based on the biomarker present. So these biomarkers will not have a condition associated with them as the biomarker is not related to disease. Instead, these biomarkers will have data for the ‘exposure_agent’ and ‘exposure_agent_ID’ fields. The 'exposure_agent' and 'exposure_agent_ID' fields are conditionally required on each other. Risk, diagnostic, prognostic, and monitoring biomarkers will have conditions associated with them instead of an exposure agent. Biomarkers may not have both condition and exposure agent data associated with them.
Biomarkers, under this definition, will have 1 condition associated with it and will have a separate entry.
Ex: condition: “prostate cancer”, condition_ID: “DOID:10283”

‘specimen’ and ‘specimen_ID’ fields are both optional in the sense that if the resource does not have the information listed, then it is acceptable to not provide these fields. These fields will be associated with the type of specimen that is used to access the biomarker. Spatial and temporal specimens can be placed in this data model as well but must follow structured ontologies and terms such as OBIB (https://pubmed.ncbi.nlm.nih.gov/27148435/). ‘Loinc_code’ is also an optional field in the same sense and is related to the lab test code associated with the biomarker from the LOINC database (LOINC Database). A biomarker can have 0, 1, or multiple LOINC codes.
Ex: specimen: “blood”, specimen_ID: “UBERON:0000178”, loinc_code: 34519-9

‘Evidence_source’ is a required field as it gives the paper(s) (as PubMed identifier; PMID) or the source(s) and identifier(s) from where the biomarker and related information was extracted. Since this field can take multiple entries for a single biomarker the approach for this would be to enter the data into an array format. ‘Evidence’ is optional as this field takes text that is related to the biomarker from the paper through manual curation or NLP extraction. The ‘notes’ field is also optional as it takes free text that can add to the contextual data for the biomarker but does not belong in any of the fields mentioned before.
Ex: evidence_source: “PMID:10914713”, evidence: “A low suPAR level (<4.75 ng/ml) at baseline is a useful biomarker for aiding clinical”

This biomarker data model is centered around the biomarker definition and also helps present the definition in a helpful and concise way. “Core” biomarker data types are the basis of the biomarker definition and will help in creating separate entries. If a “core” field is different for biomarker data this constitutes a separate entry being created. It also provides any contextual data that would be helpful in defining the biomarker. This data model is important for the project and any further additions that could help the model be further improved are always helpful and will be in consideration.

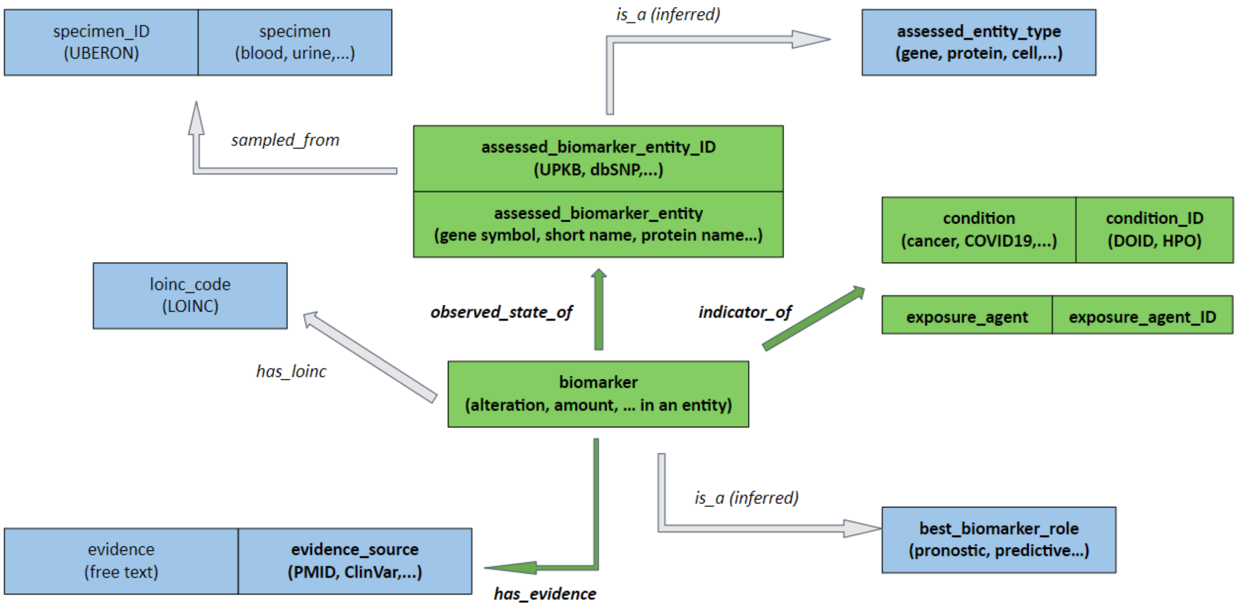

specimen_ID (UBERON)
specimen (blood, urine,...)
is_a (inferred)
assessed_entity_type (gene, protein, cell,...)
sampled_from
assessed_biomarker_entity_ID (UPKB, dbSNP,...)
assessed_biomarker_entity (gene symbol, short name, protein name...)
condition (cancer, COVID19,...)
condition_ID (DOID, HPO)
loinc_code (LOINC)
observed_state_of
indicator_of
exposure_agent
exposure_agent_ID
has_loinc
biomarker (alteration, amount, ... in an entity)
is_a (inferred)
evidence (free text)
evidence_source (PMID, ClinVar,...)
best_biomarker_role (pronostic, predictive...)
has_evidence

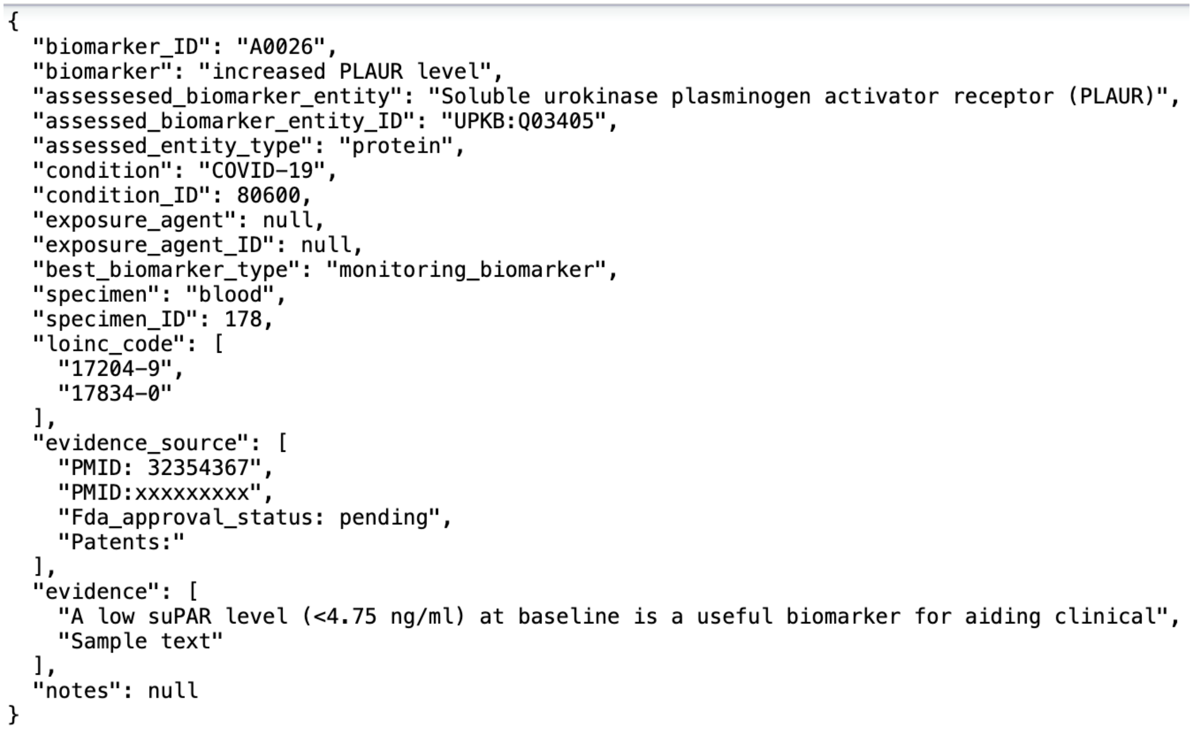
Fig 1. Biomarker Data Model. Data types and relations (named boxes and arrows, respectively) in the biomarkers of clinical interest space; “core” biomarker data types (green boxes), additional data types (blue boxes) provide valuable contextual knowledge. References and additional annotations can be added to a database implementation of the model.

Data Model JSON

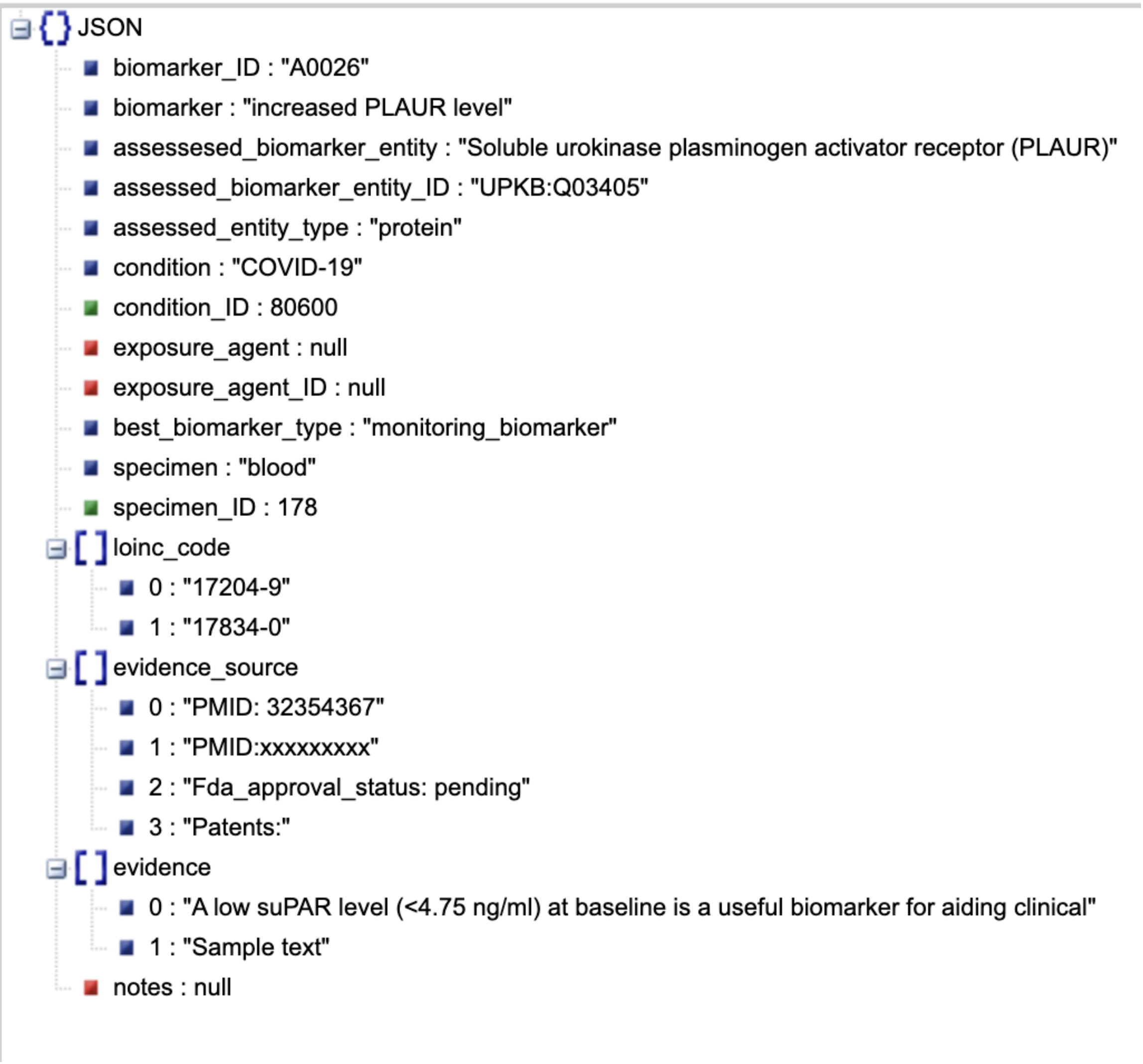
https://github.com/biomarker-ontology/biomarker-partnership/blob/main/schema/v1.1/biomarkerkb_schema.json
The JSON schema linked above has not been updated according to the comments received on the RFC yet. It will be updated later on after the finalized RFC.

The JSON schema is dynamically generated from the most current version of the data dictionary. The schema supports conditional dependencies and exclusions as well as regular expression pattern validation. The resulting schema can be used to validate incoming data before it is assigned ID’s, added to the MongoDB instance, and subsequently processed for the ontology. Once the data is validated, assigned an ID, and incorporated into the ontology, the RDF triples as assertions can be generated for the triplestore database. This will allow for the data to be represented in a knowledge graph and incorporated into other CFDE projects.

JSON Entry Example:



```
{
  "biomarker_ID": "A0026",
  "biomarker": "increased PLAUR level",
  "assessesed_biomarker_entity": "Soluble urokinase plasminogen activator receptor (PLAUR)",
  "assessed_biomarker_entity_ID": "UPKB:Q03405",
  "assessed_entity_type": "protein",
  "condition": "COVID-19",
  "condition_ID": 80600,
  "exposure_agent": null,
  "exposure_agent_ID": null,
  "best_biomarker_type": "monitoring_biomarker",
  "specimen": "blood",
  "specimen_ID": 178,
  "loinc_code": [
    "17204-9",
    "17834-0"
  ],
  "evidence_source": [
    "PMID: 32354367",
    "PMID:xxxxxxxxx",
    "Fda_approval_status: pending",
    "Patents:"
  ],
  "evidence": [
    "A low suPAR level (<4.75 ng/ml) at baseline is a useful biomarker for aiding clinical",
    "Sample text"
  ],
  "notes": null
}
```



```
⊟ { } JSON
    ■ biomarker_ID : "A0026"
    ■ biomarker : "increased PLAUR level"
    ■ assessesed_biomarker_entity : "Soluble urokinase plasminogen activator receptor (PLAUR)"
    ■ assessed_biomarker_entity_ID : "UPKB:Q03405"
    ■ assessed_entity_type : "protein"
    ■ condition : "COVID-19"
    ■ condition_ID : 80600
    ■ exposure_agent : null
    ■ exposure_agent_ID : null
    ■ best_biomarker_type : "monitoring_biomarker"
    ■ specimen : "blood"
    ■ specimen_ID : 178
  ⊟ [ ] loinc_code
        ■ 0 : "17204-9"
        ■ 1 : "17834-0"
  ⊟ [ ] evidence_source
        ■ 0 : "PMID: 32354367"
        ■ 1 : "PMID:xxxxxxxxx"
        ■ 2 : "Fda_approval_status: pending"
        ■ 3 : "Patents:"
  ⊟ [ ] evidence
        ■ 0 : "A low suPAR level (<4.75 ng/ml) at baseline is a useful biomarker for aiding clinical"
        ■ 1 : "Sample text"
    ■ notes : null
```

Proposed Actions

Biomarker Levels

A new field to add to the data model would be a field pertaining to biomarker levels. This would be a parseable element and would create a “ranking” of biomarkers based on some rules of evidence, for example by documenting the number of different evidence sources the biomarker is found in. If the biomarker is found in multiple different sources it can be given a higher score than other biomarkers found in fewer sources or just documented in a database. There are many biomarkers that have multiple studies associated with it and some biomarkers that were just examined once or twice and added to a database. We do not want to ignore any biomarkers but should have a system/set of rules to rank the biomarkers based on the number of resources they are found in. This creates a filter to only look at biomarkers with a high or low level of evidence.

Ex: “increased IL-6 levels”.This could be assigned to a certain level based on these criteria (but not limited to only these criteria):

- Primary publication with high citations
- FDA approval status
- Patent association

Multicomponent Biomarkers

The biomarker data model examples given so far have only pertained to single biomarkers. However, in preliminary data that was gathered many multicomponent/panel biomarkers have been included as well. The FNBWG proposed to define multicomponent biomarkers (MCBs) as:

- include two or more features potentially including clinical characteristics such as patient demographics;
- are used independently or in combination through an algorithm;
- represent one or more defined characteristics indicating normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions and environmental exposures.
- Further information and discussion can be found in https://www.futuremedicine.com/doi/full/10.2217/bmm-2023-0351

These could be two or more biomarker entities that exhibit a change together in a particular disease that when measured together provide a better understanding of the clinical circumstances. However, the working group also discussed many challenges and complexities of MCBs. In some cases when putting a MCB together, some components may be measured at different modalities and have more weight compared to other components in the MCB. The weight of components is determined by an algorithm set forth in the FNBWG panel discussion. The challenge for this partnership would be to handle adding MCBs within the data model. It is not impossible and one proposed way could be to add the MCB as we are doing with the single biomarkers and have all the associated fields with it. Within a JSON structure this could be easily handled and visualized as individual components of the MCB could be nested within the top level MCB with their own core fields and contextual fields filled out. To narrow the scope for MCBs, we can handle MCBs that are made of all genes, variants, metabolites, and glycans, avoiding complex biomarkers.

Multicomponent biomarkers can be handled in this data model by using an array for the biomarker component. An array will allow single biomarkers to be populated into the model with a single biomarker ID. It will also populate multiple components underneath a single biomarker if needed.