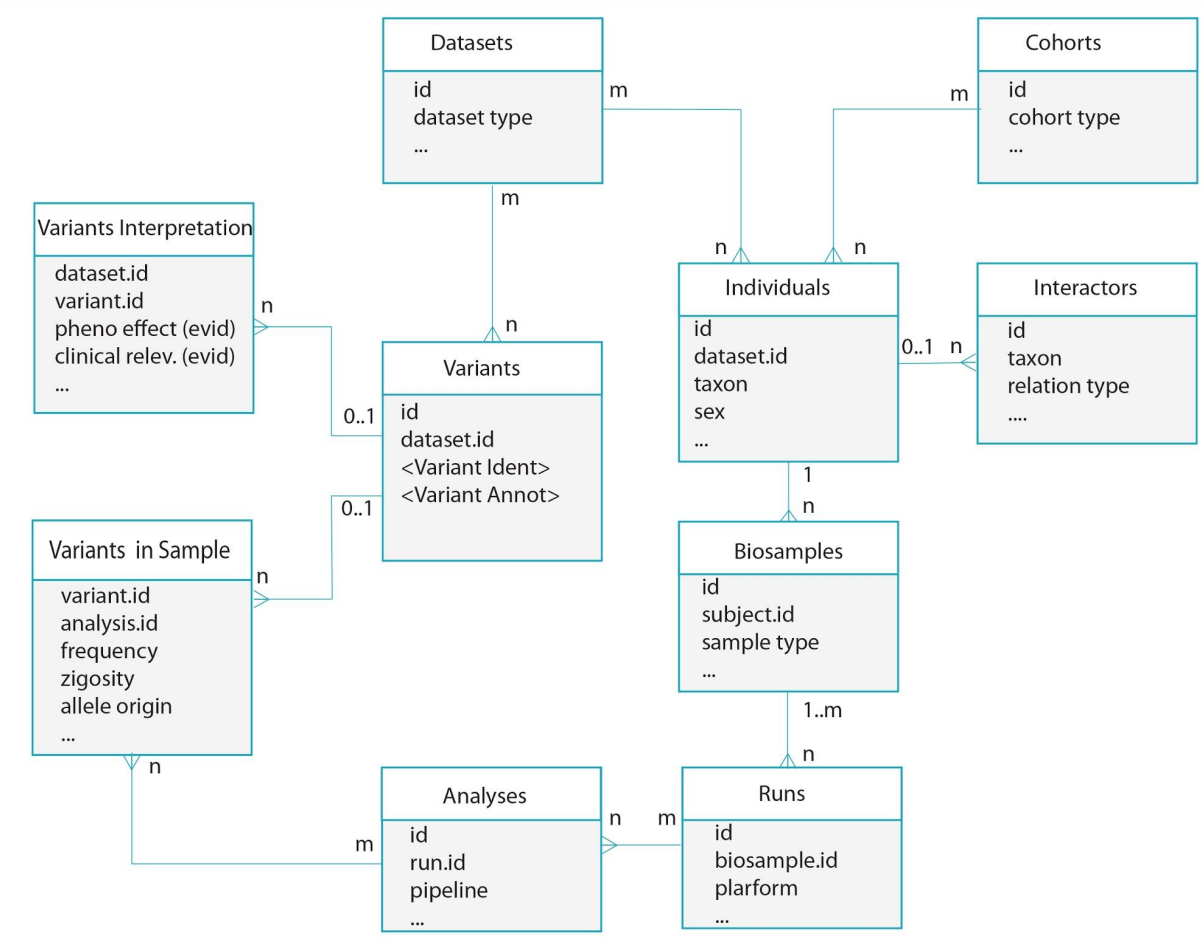
Beacon v.2 default schemas draft 3

For comments> link to this document



NOTE> Many things are moved, working on fixing this on local now
-Check the issue with Dataset object

Datasets

datasetId categorical value Dataset reference ID

datasetSource categorical value Reference to dataset source, e.g. "DECIPHER", "DisGenNET"

datasetType categorical value Type of dataset: "case-level" or "variant-level" (aggregated)

Variants Identification

variantId alphanumeric value Reference to variant ID

assemblyId categorical value Genomic assembly accession and version as RefSqeq assembly accession (e.g. "GCG 000001405.39") or a versioned assembly name or synonym such as UCSC Genome Browser assembly (e.g. "hg38") or Genome Reference consortium Human (e.g. "GRCh38.p13") names

refseqId categorical value Reference sequence Refseq ID and version for genomic reference sequence in which variant coordinates are given, e.g. "NC 000009" for human chromosome 9. Alternatively, names, synonymous or aliases e.g. "Chr9" when **assemblyId** is given. For organisms with a single reference sequence covering the genome, the versioned Refseq ID can be given here as an alternative to the assembly ID and version in **assemblyId**, e.g. "NC 045512.2" for SARS-CoV2 full-length genome reference sequence.

start numeric value Start position of variant

end numeric value End position of variant

ref alphanumeric value Reference sequence in start-end coordinates

alt alphanumeric value Alternative sequence in start-end coordinates

variantType categorical value (ontology term object) Type of variant. Value from Sequence Ontology, e.g. "SNV" (SO:0001483), "structural variant" (SO:0001537)

Variants Annotation

variantId alphanumeric value Reference to Variant ID

variantAlternativeIds (List of) alphanumeric value(s) Cross-referencing ID(s) (CURIE(s)) for the variant in the original databases or variant-level (aggregated) databases for previously described variants (e.g. clinVarId, ClinGen, COSMIC), e.g. "VCV000055583.1", "CA003602"

genomicHGVSId alphanumeric value HGVSId descriptor at genomic level (recommended, referred to genome assembly defined in Variant Identification), e.g. "NC 000017.10:g.41199678C>A"

transcriptHGVSIds (List of) alphanumeric value(s) HGVSId descriptor at transcript level: "NC 000023.10(NM 004006.2):c.357+1G

proteinHGVSIds (List of) alphanumeric value(s) HGVSId descriptor(s) at protein level (for protein-altering variants), e.g. "NP 009225.1:p.Glu1817Ter" or "LRG 199p1:p.Val25Gly" (preferred)

genomicRegions (List of) categorical value(s) (ontology term) Classification(s) of the variant according to the genomic region affected (all that apply, relative to each feature affected). Value from Ensembl Glossary (ENSGLOSS) Variant consequence ontology (ENSGLOSSARY:00000134, e.g. "3UTR" (ENSGLOSSARY:0000159), "coding sequence variant" (ENSGLOSSARY:0000159), "upstream gene variant" (ENSGLOSSARY:0000164), "intergenic variant" (ENSGLOSSARY:0000174), "intron variant" (ENSGLOSSARY:0000161), "non-coding transcript variant" (ENSGLOSSARY:0000163)

genomicFeatures Genomic feature(s) affected by the variant. (List of:)

class categorical value (ontology term object) Class of feature affected by the variant. Value from ENSGLOSS Biotype (ENSGLOSSARY:0000025) ontology, e.g. "protein coding gene", "non-coding RNA", "long non-coding RNA"

featureID (alphanumeric value) ID /accession/name of feature affected by the variant, matching **class**, e.g. "TP53", "GeneID:43740578"

molecularEffects (List of) categorical value(s) (ontology term object) Predicted effect at nucleotide level for protein affecting variants. Value from Sequence ontology (SO) coding sequence variant (SO:0001580) ontology, e.g. "synonymous variant" (SO:0001819), "nonsynonymous vari- ant" (SO:0001992) (and classifications therein, such as "stop gained" (SO:0001587), "missense variant" (SO:0001583), "inframe indel" (SO:0001820))

aminoacidChanges (List of) categorical value(s) Change at aminoacid level for protein affecting missense variants e.g. "V304*"

Individuals

individualId alphanumeric value Individual reference ID (external accession or internal ID)

datasetId alphanumeric value Reference to dataset ID

taxonId alphanumeric value) Taxon ID of Individual, e.g. "9606"

sex categorical value (ontology term) Sex of Individual. Value from NCIT General Qualifier ontol- ogy (NCIT:C27993): " UNKNOWN" (not assessed or not available) (NCIT:C17998), " FEMALE" (NCIT:C46113), " MALE" (NCIT:C46112) or " OTHER SEX" (NCIT:C45908)

ethnicity categorical value (ontology term object) Ethnic background of Individual. Value from NCIT Race ontology (NCIT:C17049). e.g. "Latin American" (NCIT:C126531)

geographicOrigin categorical value (ontology term object) Individual' s country or region of origin (birthplace or residence place regardless of ethnic origin). Value from GAZ Geographic Location ontology (GAZ:00000448), e.g. "United States of America" (GAZ:00002459)

phenotypicFeatures Phenotypic feature(s) observed in the Individual, defined by phenotype, date, type or age of onset and level/ severity. (List of:)

phenotypeId categorical value (ontology term object) Phenotypic feature observed. Value from Human Phenotype Ontology (HPO)

dateOfOnset alphanumeric value (ISO8601 date format) Date of onset/observation of phenotype, e.g. "2010-07-10"

onsetType categorical value (ontology term object) Onset type. Value from HPO Onset ontology (HP:0003674), e.g. "congenital onset" (HP:0003577), "adult onset" (HP:0003581)

ageOfOnset Individual' s age at onset/observation of phenotype

age alphanumeric value (ISO8601 duration format) Age, e.g. "P32Y6M" **ageGroup** categorical value (ontology term object) Age group. Value from NCIT Age Group ontology, e.g. " NCIT:C27954" (Adolescent)

level/severity categorical value (ontology term object) Level/severity when and as applicable to phenotype observed. Value from Human Phenotype Ontology (HPO) Severity ontology (HP:0012824), e.g. "severe" (HP:0012828)

diseases Disease(s) been diagnosed to the Individual, defined by disease ID, date, type or age of onset, stage, level/severity and the presence of family history. (List of:)

diseaseId categorical value (disease code /ontology term object) Disease ID. Value from ICD10 disease codes or ontology terms from disease ontologies such as HPO, OMIM, Orphanet, MONDO, e.g. "lactose intolerance" (HP:0004789)

dateOfOnset alphanumeric value (ISO8601 date format) Date of onset/diagnosis of disease, e.g. "1987-10-28"

onsetType categorical value (ontology term object) Onset type. Value from HPO Onset ontology (HP:0003674), e.g. "congenital onset" (HP:0003577), "adult onset" (HP:0003581)

ageOfOnset Individual' s age at onset/ diagnosis of disease

age alphanumeric value (ISO8601 duration format) Age, e.g. "P6M" **ageGroup** categorical value (ontology term object) Age group. Value from NCIT Age Group ontology, e.g. "NCIT:C27954" (Adolescent)

stage categorical value (ontology term object) Stage of disease. Value from Ontology for General Medical Science or Disease Stage Qualifier ontology (NCIT:C28108), e.g. " acute onset" (OGMS:0000119)

level/severity categorical value (ontology term object) Level/severity when and as appli- cable to disease course. Value from Human Phenotype Ontology (HPO) Severity ontology (HP:0012824), e.g. "mild" (HP:0012825)

familyHistory boolean indicating determined or self-reported presence of family history of the disease

treatments Treatment(s) or Medication(s) been prescribed/administered to Individual, defined by treat ment ID, date and age of onset, dose, schedule and duration. (List of:)

treatmentId categorical value (ontology term object) Treatment ID. Value from NCIT Drug, Food, Chemical or Biomedical Material.

treatmentRoute categorical value (ontology term object) Route of treatment. Value from NCIT Route of Administration ontology.

dateAtOnset alphanumeric value (ISO8601 duration format) Date of the beginning of treatment

ageAtOnset Individual' s age at the beginning of treatment

age alphanumeric value (ISO8601 duration format) Age

ageGroup categorical value (ontology term object) Age group value, from NCIT Age Group ontology, e.g. "NCIT:C27954" (Adolescent)

dose numeric Treatment dose

units alphanumeric Treatment dose units

schedule categorical value (ontology term object) Treatment schedule. Value from NCIT Schedule Frequency ontology, e.g "weekly" (NCIT:C67069)

duration alphanumeric value (ISO8601 duration format) Treatment duration

interventions Intervention(s) been practiced on Individual, defined by intervention ID), date and age of onset, dose, schedule and duration. (List of:)

interventionId categorical value (ontology term object) Intervention ID. Value from Medical Action Ontology.

modifier Modifier of the intervention. Value from TBD.

date alphanumeric value (ISO8601 duration format) Date of intervention

ageAtIntervention Individual' s age at the date of intervention, as age or age range

age alphanumeric value (ISO8601 duration format) Age

ageGroup categorical value (ontology term object) Age group value, from NCIT Age Group ontology, e.g. "NCIT:C27954" (Adolescent)

measures Measures(s) been taken from Individual, defined by measure ID, date and/or age at measure ment, unit and value. (List of:)

measureId categorical value (ontology term object) Measure ID. Value from Clinical Mea- surement ontology (CMO)

modifier Modifier of the measurement. Value from TBD.

date alphanumeric value (ISO8601 duration format) Date of intervention

ageAtMeasure Individual' s age at the date of measurement, as age or age range

age alphanumeric value (ISO8601 duration format) Age

ageGroup categorical value (ontology term object) Age group value, from NCIT Age Group ontology, e.g. "NCIT:C27954" (Adolescent)

exposures Exposures(s) occurred to or practiced (lifestyle, behavioural exposures) by Individual, defined by exposure ID, date and age of onset, dose, schedule and duration. (List of:)

interventionId categorical value (ontology term object) Intervention ID. Value from En- vironment Exposure Ontology.

modifier Modifier of the exposure. Value from TBD.

date alphanumeric value (ISO8601 duration format) Date of intervention

ageAtIntervention Individual' s age at the date of intervention, as age or age range

age alphanumeric value (ISO8601 duration format) Age

ageGroup categorical value (ontology term object) Age group value, from NCIT Age Group ontology, e.g. "NCIT:C27954" (Adolescent)

pedigrees Pedigree(s) to which the Individual belongs. (List of:)

pedigreeId alphanumeric value Reference to pedigree ID (external accession)

pedigreeDisease categorical value (disease code/ontology term object) Disease ID. Value from ICD10 disease codes or ontology terms from disease ontologies such as HPO, OMIM, Orphanet, MONDO, e.g. "lactose intolerance" (HP:0004789)

pedigreeRole categorical value (ontology term object) Pedigree role, defined as relation- ship to proband. Value from Family member ontology (NCIT:C41256), e.g. "proband" (NCIT:C64435), "identical twin" (NCIT:C73429), " mother" (NCIT:C25189)

affectedStatus categorical value Affected status of Individual in disease of pedigree: "af- fected" or

"unaffected"

numSubjects numeric value Number of subjects in pedigree

Biosamples

biosampleId alphanumeric value Biosample reference ID (external accession)

subjectId Reference to Individual ID

description free text Any relevant info about the biosample that does not fit in any field in the schema

biosampleStatus categorical value (ontology term object) Classification of biosample based on their

role in study. Value from Experimental Factor Ontology (EFO) Material Sample ontology (OBI:0000747) Classification of the sample in " abnormal sample " (EFO:0009655) or " reference sample" (EFO:0009654)

collectionDate alphanumeric value (ISO8601 date format) Date of biosample collection, e.g. "2019-12"

subjectAgeAtCollection alphanumeric value (ISO8601 duration format) Individual' s age at the time of biosample collection, e.g. "P72Y6M19D"

sampleOriginDescriptors Descriptor(s) of origin of biosample. (List of:)

sampleOriginType categorical value (ontology term object) Category of sample origin. Value from Ontology for Biomedical Investigations (OBI) material entity (BFO:0000040) on- tology, e.g. "specimen from organism" (OBI:0001479), "xenograft" (OBI:0100058), "cell cul- ture" (OBI:0001876), "cell specimen" (OBI:0001468), "environmental swab specimen" (OBI:0002613)

sampleOriginDetail categorical value (ontology term object) Specific instance of sample origin matching the category set in **sampleOriginType**. Value from Uber-anatomy ontology (UBERON) or BRENDA tissue / enzyme source (BTO), Ontology for Biomedical Investi- gations (OBI) or Cell Line Ontology (CLO), e.g. "cerebellar vermis" (UBERON:0004720), "HEK-293T cell" (BTO:0002181), "nasopharyngeal swab specimen" (OBI:0002606), "cere- brospinal fluid specimen" (OBI:0002502)

obtentionProcedure categorical value (ontology term object) Ontology ID from NCI Thesaurus (NCIT) Intervention or Procedure (NCIT:C25218) ontology. e.g. "biopsy" (NCIT:C15189)

cancerFeatures Values specifying cancer-specific features, including progression and tumor grade

tumorProgression categorical value (ontology term object). Descriptor of tumor progres- sion. Value from Neoplasm by Special Category ontology (NCIT:C7062). Tumor progression category indicating primary, metastatic or recurrent progression e.g. "Primary Malignant Neo- plasm" (NCIT:C84509)

tumorGrade categorical value (ontology ID) Tumor grading. Value from Tumor Grading Characteristic ontology (Mondo Disease Ontology MONDO:0024488), e.g. "Grade 3" (MONDO:0024493)

Runs

runId alphanumeric value Run reference ID (external accession or internal ID)

biosampleId alphanumeric value Reference to Biosample ID

runDate alphanumeric value (ISO8601 date format) Date at which run was performed , e.g. "2019-12-28"

librarySource categorical value Sequencing library source, e.g. "Metagenomic" , "Viral RNA"

libraryStrategy categorical value Sequencing library strategy, e.g. "WGS"

librarySelection categorical value Selection method for sequencing library preparation, e.g. "RAN- DOM" , "RT-PCR"

libraryLayout categorical value Sequencing library layout, e.g. "PAIRED" , "SINGLE"

platform categorical value Sequencing technology, e.g. “Illumina”, “Oxford Nanopore Technologies”

platformModel categorical value Sequencing platform model, e.g. “Illumina MiSeq”, “GridION”

Analyses

analysisId alphanumeric value Analysis reference ID (external accession or internal ID)

runId alphanumeric value Reference to run ID

analysisDate alphanumeric value (ISO8601 date format) Date at which analysis was performed, e.g. “2020-01-10”

pipelineName categorical value Analysis pipeline and version, e.g., “Isaac Whole Genome Sequencing; version 3.0.0.0”

pipelineRef Link to Analysis pipeline resource

aligner categorical value Mapping/Alignment software e.g. “bwa-0.7.8”

variantCaller categorical value Variant calling software/ pipeline, e.g. “GATK4.0”

Variants in Sample

variantId alphanumeric value Reference to variant ID

analysisId Reference to analysis ID

biosampleId Reference to biosample ID

variantFrequency numeric value Variant/allele frequency in biosample

zigosity categorical value (ontology term object) Zigosity in which variant is present in the sample.

Value from the Zigosity Ontology (GENO:0000133), e.g. “heterozygous” (GENO:0000135)

alleleOrigin categorical value (ontology term object) Allele origin of variant in sample. Value from the Variant Origin (SO:0001762), e.g. “somatic variant”, “germline variant”, “de novo variant”.

phenotypicEffects Observed effect(s) of variant on phenotype. (List of:)

phenotypeId categorical value (ontology term object) Descriptor of phenotype found as- sociated to variant in the present study. Value from Human Phenotype Ontology (HPO)

phenotypeEffect categorical value (ontology term object) Phenotypic effect classification determined in the present study. Value from Sequence types and features ontology (SO) variant phenotype (SO:0001769), e.g. “quantitative variant” (SO:0001774)

evidenceType categorical value (ontology term object) Type of evidence supporting variant-phenotype association from the Evidence & Conclusion Ontology (ECO), e.g. “experimental evidence”

clinicalRelevances Observed effect of variant on disease. (List of:)

diseaseId categorical value (disease code/ ontology term object) Descriptor of disease associated. Value from ICD10 disease codes or ontology terms from disease ontologies such as HPO, OMIM, Orphanet, MONDO, e.g. “lactose intolerance” (HP:0004789)

clinicalEffect categorical value (ontology term) Clinical effect classification. Value from Sequence types and features ontology (SO) variant phenotype (SO:0001769), e.g. “disease causing variant” (SO:0001772)

evidenceType categorical value (ontology term object) Type of evidence supporting variant- disease association from the Evidence & Conclusion Ontology (ECO), e.g. “experimental evidence”

Variants Interpretation

variantId alphanumeric value Reference to Variant ID

datasetId alphanumeric value Reference to dataset ID source of variant interpretation

phenotypicEffects Annotated effect(s) of variant on any phenotypic feature other than a disease. (List of:)

> **phenotypeId** categorical value (ontology term object) Descriptor of phenotype associated. Value from Human Phenotype Ontology (HPO)
>
> **phenotypeEffect** categorical value (ontology term object) Phenotypic effect classification. Value from Sequence types and features ontology (SO) variant phenotype (SO:0001769), e.g. "benign variant" (SO:0001770)
>
> **alleleOrigin** (List of) categorical value(s) (ontology term object) Annotation(s) on allele origins in which the variant has been found in association to phenotype. Value(s) from Se- quence types and features ontology (SO) variant origin (SO:0001762) ontology, e.g. "somatic variant" (SO:0001777), "germline variant" (SO:0001778), "de novo variant" (SO:0001781)
>
> **references** (List of) publication identifier(s) Reference(s) to publication(s) describing association, as PMID(s), e.g. "PMID:17024373"

clinicalRelevances Annotated effect(s) of variant on disease. (List of:)

> **diseaseId** categorical value (disease code/ ontology term object) Descriptor of disease associated. Value from ICD10 disease codes or ontology terms from disease ontologies such as HPO, OMIM, Orphanet, MONDO. e.g. "lactose intolerance" (ICD10CM:E73)
>
> **clinicalEffect** categorical value (ontology term object) Clinical effect classification. Value from Sequence types and features ontology (SO) variant phenotype (SO:0001769), e.g. "dis- ease associated variant" (SO:0001771)
>
> **alleleOrigin** (List of) categorical value(s) (ontology term object) Annotation(s) on allele origins in which the variant has been in association to condition. Value(s) from Sequence types and features ontology (SO) variant origin (SO:0001762) ontology, e.g. "somatic variant" (SO:0001777), "germline variant" (SO:0001778), "de novo variant" (SO:0001781)
>
> **references** (List of) publication identifier(s) Reference(s) to publication(s) describing association, as PMID(s), e.g. "PMID:19793655"

Interactors

This is an organism/agent whose phenotypic data is collected in association with the Individual, but which is not the sequenced/genotyped Individual itself. This block accounts for the ' extended phenotype' of variants in organisms/agents other than the one harboring them.

relationType categorical value (ontology term object) Type of relation with Individual. Value from Infectious disease Ontology(IDO), e.g. "host" (IDO:0000531), "commensal" (IDO:0000525), "in- fectious agent" (IDO:0000596)

[...] All the rest of objects from Individual

Cohorts (link)

cohortType categorical value Cohort type by its definition. One of: ´study-defined´, ´beacon-defined´ or ´user-defined´. If a cohort is declared ´study-defined´ or ´beacon-defined´, criteria are to be entered in **cohortInclusionCriteria**; if a cohort is declared 'user-defined' **cohortInclusionCriteria** will be automatically populated from the parameters used to perform the query.

cohortId ALPHANUMERIC Cohort identifier. For ´study-defined´ or ´beacon-defined´cohorts this field is set by the implementer. For ´user-defined´ this unique identifier could be generated upon the query that defined the cohort, but could be later edited by the user

cohortName ALPHANUMERIC Cohort name. For ´user-defined´ this field could be generated upon the query, e.g. a

value that is a concatenation or some representation of the user query

cohortInclusionCriteria Inclusion criteria. For ´beacon-defined´cohorts, cohorts matching the whole dataset will not apply criteria for if cohorts are subsets of Datasets the criteria used to defined them can be added in these fields). (List of:)

cohortLocations categorical value Cohort type by its design

cohortGenders categorical value Cohort type by its design

cohortEthnicities categorical value Cohort type by its design

cohortDiseases categorical value Cohort type by its design

cohortPhenotypicFeatures categorical value Cohort type by its design

cohortAgeRange categorical value Individual age range in cohort inclusion criteria

start alphanumeric value (ISO8601 duration format) Min age accepted in cohort criteria

end alphanumeric value (ISO8601 duration format) Max age accepted in cohort criteria

cohortExclusionCriteria Exclusion criteria. Same structure as **cohortInclusionCriteria**

cohortLicense TBD

cohortRight TBD

cohortContact TBD

cohortDataTypes Data types available from the Cohort. (List of:)

dataType categorical value (ontology term object) Data types included in cohort (GECKO, e.g. Genomic data (GECKO:0000032), Clinical Measurements (CMO:0000000), Lifestyle (GECKO:0000067), etc.
num paired to id field, number of individuals with this data type in the cohort

collection_events Collection events/ data points (List of:)

collection_event

event_num integer Cardinality of the collection event/data point in a series

event_date ISO8601 date and time format date of collection event/data point.

event_timeline Aggregated information of dates of visit/diagnostic/inclusion in study obtained from individual level info in database. Will coincide with collection event date for multi-time

start alphanumeric value (ISO8601 date format) Earliest date of event

end alphanumeric value (ISO8601 date format) Latest date of event

event_size integer Count of individuals in cohort at data point (for ´user-defined´ cohorts, this is individuals meeting criteria) obtained from individual level info in database.

event_cases integer count of cases

event_controls integer Count of controls

event_locations Aggregated information of geographic location obtained from individual level info in database.

availability boolean availability

availability_count integer count of individuals with data available

distribution list of: [unique values in field, count]

event_genders (Aggregated information of gender(s) obtained from individual level info in database.

availability boolean data availability

availability_count integer count of individuals with data available

distribution (List of) [unique values in field, count]

event_ethnicities Aggregated information of ethnicity obtained from individual level info in database.

availability boolean data availability

availability_count integer count of individuals with data available

distribution (List of) [unique values in field, count]

event_age_range Individual age range, obtained from individual level info in database

start: alphanumeric value (ISO8601 duration format) Min age in collection event

end alphanumeric value (ISO8601 duration format) Max age in collection event

event_diseases Aggregated information of disease/condition(s) obtained from individual level info in database.

availability boolean data availability

availability_count integer count of individuals with data available

distribution (List of) [unique values in field, count]

event_phenotypes (GECKO disease or disorder: MONDO:0000001) Aggregated information of phenotype(s) obtained from individual level info in database.

availability boolean data availability

availability_count integer count of individuals with data available

distribution (List of) [unique values in field, count]

event_data_types (list of) Aggregated data type information available for each cohort data type as declared in cohort_data types, and obtained from individual level info

availability boolean data availability

availability_count integer count of individuals with data available

distribution (List of) [unique values in field, count]