ClinVar Data Dictionary

Overview

This document defines the data elements represented in the ClinVar database. The document includes descriptions of how data are managed, the XML used to represent each concept for submission (see ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_submission.xsd), the field name in the spreadsheet version of the submission document, the table and column in which the data are stored in the relational database, and allowed values. Please note, the xml used for export (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xml/) is slightly different from that used for submission, and is validated by a distinct xsd (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/xsd_public/)

The database has a flexible data model, so submissions may be minimal or very detailed. The focus of any submission can be considered at multiple levels, namely one aimed at more general variant-level data and the other for more specific case-level data and supporting evidence.

Standard terms used by ClinVar (and coordinated with the NIH Genetic Testing Registry (GTR)), are summarized here: http://www.ncbi.nlm.nih.gov/clinvar/docs/authorities/

Status of this document

PublicRelease2. June 30, 2014. Please direct any comments to clinvar@ncbi.nlm.nih.gov

General processing

Data added computationally

Not all values included in this document are supplied from a submitter; some are added based on information in NCBI's databases. These values are marked explicitly as '**from NCBI**'.

Optional and required values

Some elements are hierarchical in representation. If a major category topic is optional, all data elements in that category are optional. But if an optional category is selected, then the data elements listed as required are required for that category.

Validation of submissions

Enforcing the rules presented in this document is not managed only via the xsd provided from our ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_submission.xsd). Some validation is provided at the database level, by comparison to standard terminologies or known relationships among variants, genes, and conditions, or by validating reported alleles against the stated reference sequence. Standard terms used by ClinVar (and coordinated with the NIH Genetic

Testing Registry (GTR)), are summarized here:
http://www.ncbi.nlm.nih.gov/clinvar/docs/authorities/

Conventions used in this document

The following elements are provided in the sections that detail each data element:

- definition (text)
- requirements (format, optional)
- Representation on submission forms (spreadsheet, XML)
- Representation in the relational component of the database

Data representations used in multiple contexts

Source/Status

Most elements in the database are characterized with respect to the source of the information, identifiers used by the submitter or other data sources, date submitted, date modified, status of the record (*e.g.* current/to be deleted/secondary to another record), review status, and whether the data should be public or private. Rather than repeating these elements for each data category defined below, the word **Source/Status** is used as a pointer to the Data source/Status section, where the source and status elements are defined.

AttributeSet

Many concepts in the database are represented by what ClinVar terms an AttributeSet, which is an open-ended structure providing the equivalent of a type of information, the value(s) for that data type, submitter(s), free text comment(s) describing that attribute, identifier(s) for that attribute, and citation(s) related to that attribute. Rather than repeating this description per attribute, the word **AttributeSet** is used to indicate that the data are stored using this data structure, with the attribute types expected for that database concept. Thus, by definition, an **AttributeSet** includes **Source/Status**.

Data source

ClinVar maintains attribution for each data element based on the description of the person and organization providing the information. In the database, these concepts are maintained by identifiers for the organization and identifiers for the individual.

XML: XRef
db: GTR.clinvar.attr_source.extrn_src (organization)
db: GTR.clinvar.attr_source.entered_by (individual)

Identifiers in public database records (optional)

The database cross-reference structure (XRef) is provided to represent pointers to identifiers in other databases for the same concept. For example, if a gene is being described, then XRefs can be provided to NCBI's Gene database, Ensembl, HGNC*, etc.*, including the database, the identifier, and URL.

Spreadsheet: multiple locations
XML: XRef/@db, XRef/@id, XRef/@url (DB, ID in the public XML)
db: GTR.clinvar.attr_source

Citations (optional)

Citations include published articles and URLs. If a database name and identifier are supplied, the full text is not required.

Type

NCBI classifies citations by type, based in part on curation but also on the publication type provided by PubMed. These types are reported in our public XML, but not expected to be provided by submitters. Values include general, review, practice guideline, Position Statement, Translational/Evidence-based, Suggested Reading, and Recommendation.

The spreadsheet has the following citation columns:

Column	Tab	Purpose
Clinical significance citations	Variant	Citations documenting the clinical significance (by ID)
Citations or URLs for clinical significance without database identifiers	Variant	Text or URL citations documenting the clinical significance
Evidence citations	CaseData	Citations documenting the evidence (by ID)
Citations or URLs for evidence without database identifiers	CaseData	Text or URL citations documenting the evidence
Method citations	CaseData	Citations documenting the method (by ID)
Citation	SubmissionInfo	Citations applicable to the complete submission set.

Spreadsheet: All citations default to citation type “general”
XML: /Citation @Type
db: GTR.dbo.citation.citation_type

Source

The name of the data service providing an identifier for a citation. This value should not be completed if the citation is a URL or free text. Current options include PubMed, PubMedCentral, DOI, NCBI bookshelf

Spreadsheet: Citation (Source default = PubMed)
XML: Citation/ID@Source
db: GTR.dbo.citation.extrn.src

ID: the identifier provided by that data source for a citation. This value should not be completed if the citation is a URL or a free text.

Spreadsheet: Citation
XML: Citation/ID
db: GTR.dbo.citation.extrn_id

URL: complete URL

Spreadsheet: Citations or URLs without database identifiers
XML: Citation/URL
db: GTR.dbo.citation.url

CitationText: This should be used only when there is no database ID for the publication and no URL.

Spreadsheet: Citations or URLs without database identifiers
XML: Citation/CitationText
db: GTR.dbo.citation.citation

Comments (optional)

A free text comment can be provided to describe submitted data. Line breaks are retained, but no other formating. In the database, each comment is connected to the content about which a comment is made, based on the name of the database table and the unique identifier in that table. This database (db) implementation is not documented in each section where comments are supported.

Text: (required) the content of the comment

Spreadsheet: multiple locations (comment or private comment)
XML: Comment/CommentText
db: GTR.dbo.comment.comment

Type: (required) public (is rendered on the web) or private (for internal use; to explain a submission and be stored in the database but not rendered on the web) .

Spreadsheet: multiple locations (comment or private comment)
XML: Comment/Type
db: GTR.dbo.comment.comment_type

Information describing the submitter and the submission

Identification of the submitter

ClinVar uses multiple methods to identify a submitter. One type is the person to contact, listed as 'SubmitterOfRecord' in the XML. The other is the optional official submitter, if the official submitter is different from the contact.

Submitter of record (required)/other official submitters (optional)

Submitter name (choice with SubmitterID/SubmitterIDType)

The first and last name of the person actually submitting this batch. If you know your identifier in dbSNP (submitter handle) or GTR/ClinVar (Person ID), you can identify the submitter of record by those identifiers.

Spreadsheet: SubmissionInfo.Submitter first name
Spreadsheet:SubmissionInfo.Submitter last name
XML: //Person/Name/First
XML: //Person/Name/Last
db: GTR.dbo.person

Submitter Identifier

NCBI maintains several identifier systems for submitters (*e.g.* the dbSNP submitter handle), and there may be public identifier systems as well. Thus the identifier for a submitter is managed as a database cross-reference. The submitter handle is treated explicitly in the database.

Spreadsheet: SubmissionInfo.SubmitterID
Spreadsheet: SubmissionInfo.SubmitterIDType
XML: SubmitterOfRecord/Person/SubmitterHandle
XML: Submitter.Personnel.PersonRef/@id and @db where db=snp
db: GTR.dbo.person.submitter_handle
db: GTR.dbo.person.submitter_id

Private contact information (required)

The contact information of the person actually making the submission. This is required, but not publicly displayed; it is used for contact regarding submissions only.

Spreadsheet: SubmissionInfo.Submitter type
Spreadsheet: SubmissionInfo.Submitter email

Spreadsheet: SubmissionInfo.Submitter phone
XML: Submitter/Personnel/PrivateContact/email
XML: Submitter/Personnel/PrivateContact/phone
XML: Submitter/Personnel/PrivateContact/fax
db: GTR.dbo.contact (identified as private in GTR.dbo.org_person)

Public Contact information (optional, publicly displayed)

Spreadsheet: SubmissionInfo.Submitter type
Spreadsheet: SubmissionInfo.Submitter email
Spreadsheet: SubmissionInfo.Submitter phone
XML: Submitter/Personnel/PublicContact/email
XML: Submitter/Personnel/PublicContact/phone
XML: Submitter/Personnel/PublicContact/fax
db: GTR.dbo.contact (identified as public in GTR.dbo.org_person)

Organization (required)

The organization responsible for the submission

Spreadsheet: SubmissionInfo.Organization
Spreadsheet: SubmissionInfo.OrganizationID
Spreadsheet: SubmissionInfo.Institution
Spreadsheet: SubmissionInfo.Street
Spreadsheet: SubmissionInfo.City
Spreadsheet: SubmissionInfo.State/Province
Spreadsheet: SubmissionInfo.Country
Spreadsheet: SubmissionInfo.Postal code
XML: SubmitterOfRecord/Organization/Name
XML: SubmitterOfRecord/Organization/NCBIOrganizationID
XML: Submitter/Organization/Institution
XML: Submitter/Organization/StreetAddress/Line1
XML: Submitter/Organization/StreetAddress/City
XML: Submitter/Organization/StreetAddress/State
XML: Submitter.Organization/StreetAddress/PostCode
db: GTR.dbo.contact (identified as private in GTR.dbo.org_person)

Descriptors of the submission

Date submitted

- required

This may be provided explicitly. If not provided, then the date a submission is processed is used as the submission date. If the submission is an update of an existing record, the submission date is the date of record of a new version of the submission.
Spreadsheet: SubmitterInfo.Submission date

XML: ClinvarSubmissionSet/@Date
db: GTR.clinvar.measure_target.subdate

Review status

- optional

Review status indicates the level of confidence in any assertion. The values to be selected depend in part on the type of submission (in parentheses after the values listed below). Values are:

- not classified by submitter (SCV, RCV from NCBI)
- classified by single submitter (SCV, RCV from NCBI)
- reviewed by expert panel (SCV, RCV from NCBI)
- reviewed by professional society (RCV) (includes practice guidelines)
- conflict identified (RCV, from NCBI)

If no value is provided, this defaults to "classified by single submitter" except where clinical significance is "not provided," in which case it defaults to "not classified by submitter."

Spreadsheet: SubmissionInfo.Review status
XML: MeasureTrait/ClinicalSignificance/ReviewStatus
db: GTR.clinvar.measure_target.rev_stat

Content note:

A submitter may not self-identify as an expert panel. Please review this document for the application form.

http://www.ncbi.nlm.nih.gov/clinvar/docs/expert_panel/

Release status

- required

This field supports managing a submission with a temporary hold on data being presented publicly. Allowed values are public or hold until published. If not supplied, public is the default. Submitters wishing to obtain accessions pre-publication should select 'hold until published'.
Spreadsheet: SubmissionInfo.Release status
XML: ClinvarSubmissionSet/ReleaseStatus
db: GTR.clinvar.measure_target.pubstat (record level submission)
db: GTR.clinvar.mset.pubstat

Study name

- optional

Public name of a study submitting these data and providing the sample. Can be used to indicate the name of a study population or cohort. Example: Framingham.

Spreadsheet: SubmissionInfo.Study name
XML: ClinvarSubmissionSet/StudyName
db: GTR.clinvar.mt_set_attr where attr_type = 731

Study description

- optional

Description of the study generating the submission.

Spreadsheet: SubmissionInfo.Study description
XML: ClinvarSubmissionSet/Comment
db: GTR.dbo.comment

Submission name

- optional

Use this field to identify your submission. If you supply one, we index that value so your records can be retrieved from the ClinVar interface by that name.

Spreadsheet: SubmissionInfo. Display lab name on attribution list
XML: //ClinvarSubmissionSet/@set_key
db: GTR.clinvar.mt_set.set_key

Attribution

- optional

Indicate whether your lab name should be displayed on the submitter attribution page with counts of submissions. If your answer is no, your group will not be acknowledged publicly as a submitter for any submission. This is maintained as an attribute of an organization, not of a submission.

Spreadsheet: SubmissionInfo. Display lab name on attribution list
XML: TBD
db: GTR.db.org

Indicate whether your lab name should be displayed with submitted assertions of clinical significance. "No" means your assertions will be submitted anonymously with a randomly assigned lab name, e.g. Laboratory 12345. This is maintained as an attribute of an organization, not of a submission.

Spreadsheet: SubmissionInfo. Display lab name on assertions
XML: TBD

db: GTR.dbo.org_attr

Accessions and versions

SCV

The accession for each submitted assertion. This is provided by NCBI, but is required in a submission that is an update. Submitters can request accessions be assigned in advance of a submission, and then include them as part of the submission.

Spreadsheet: Variant.ClinVarAccession
XML: ClinvarSubmissionAcc /@Acc
XML: ClinvarSubmissionAcc/ClinvarSubmissionAccType/[@val_type="name"] SCV
db: GTR.clinvar.measure_target.accession

RCV

The accession for a reviewed assertion, provided by NCBI after aggregating data from multiple submissions, or assigned to submissions from professional societies. This accession should **not** be included in submissions.

Spreadsheet: Variant.ClinVarAccession
XML: ClinvarSubmissionAcc /@Acc
XML: ClinvarSubmissionAcc/ClinvarSubmissionAccType/[@val_type="name"] RCV
db: GTR.clinvar.measure_target.accession

Record Status

- optional

If ClinVar accessions are included in a submission, record status indicates if the submission is novel (and accessions were reserved prior to submission), an update to existing SCV records, or to delete an existing SCV record.

Spreadsheet: Variant.Novel or Update
XML: ClinvarSubmission/RecordStatus
XML: ClinvarSubmissionAcc/ClinvarSubmissionAccType/[@val_type="name"] RCV
db: GTR.clinvar.measure_target.accession

Replaces ClinVarAccessions

Optional. For updates in which one or more ClinVar submitted records (SCVs) are being merged into another existing ClinVar submitted record.

Spreadsheet: Variant.Replaces ClinVarAccessions
XML: ClinvarSubmission/ReplacesAccession
db: GTR.dbo.id_hist

Condition

Information about condition is represented by the combination of type of term, the term value, and the relationship of that term to other terms in the submission. The condition information must be connected at the variant level, but can also be represented as part of each set of observations.

When the condition is defined as part of the condition-variant relationship, that condition should be the one about which the clinical assessment is being made. For phenotypic measures *observed* in one more more patients, those data should be provided as Clinical features.

Please note that ClinVar retains submitted terms, but maps these to controlled terms whenever possible. To facilitate that mapping, we encourage submitters to provide their condition descriptions as the combination of a database name and a database identifier, rather than free text.

ClinVar requires categorization of condition information. This may be done at the level of a set of conditions, a relationship between conditions, or a single condition. If the combination of term and type provided in the submission is not consistent with ClinVar's representation, ClinVar may re-assign the condition category. Current options for condition categories are:

Controlled term	Usage	Data elements
Disease	Use for a diagnostic name.	TraitSetType TraitRelationship TraitType
Drug response	Usually written as drug name + response. This includes pharmacodynamic and pharmacokinetic differences.	TraitSetType TraitRelationship TraitType
Subphenotype	Use to submit a disease hierarchy.	TraitRelationship TraitType
Blood group	For the name of a blood group system. If an allele of a blood group is manifested as an additional phenotype, include in the trait set.	TraitRelationship TraitType
Finding	Use for clinical features or phenotypic measures.	TraitRelationship TraitType
Infection resistance	Corresponds to "genetic resistance to infectious agent", **IDO_0000587**	TraitRelationship TraitType

Sets of conditions

Category/type of condition set

More than one term may be used to describe condition, e.g. a set of clinical findings in addition to a diagnosis, or multiple disease terms if it is the combination of diseases about which an assertion is being made. Each condition term is captured explicitly [see the next section, Description of one condition (trait)], and the co-occurring conditions are represented as a set. Options for type of condition sets are Disease, Drug Response, and Finding.

Content note (should condition names be combined?)

To report that a variant is pathogenic for disease1 and also pathogenic for disease2 submit the data on multiple rows; one set specific to the variant-disease1 combination and one specific to variant-disease2. If, on the other hand, to report that a variant is pathogenic only in the context of the combination of disease1+disease2, submit on the same row in the spreadsheet or in the same TraitSet in the XML.

Content note (no diagnostic term, but observed features)

When the name of a condition is not known, but multiple clinical features are being reported at the case level, please submit Preferred condition name as 'see cases'. 'See cases' will then provide users of the data that a diagnosis has not been made, but the clinical features are provided.

Spreadsheet: Variant.Condition category
XML: TraitSet/TraitSetType[@val_type="name"]
db: GTR.clinvar.tset.type

Relationships among members of a set of conditions

The relationship among a set of conditions may be described. Some examples:

- a set of a clinical features and a diagnostic name may be represented as several Findings and a Disease.
- sickle cell anemia and resistance to malaria may be represented as Disease and Infection resistance.
- subphenotype may be used for hierarchical relationships, for example Usher Syndrome, type 1B may be represented as a subphenotype of Usher Syndrome, type 1. The ClinVar/GTR staff curates some hierarchical relationships, but usually uses those provided by external authorities. If you wish to suggest a revision of current hierarchies, or suggest new ones, please contact us. Note that severity should be represented as an attribute of the phenotype (see below), rather than as a subphenotype.

Spreadsheet: not represented
XML: Trait/TraitRelationship/TraitRelationshipType/@val_type = "name"

db: GTR.clinvar.tsubset.relat_type

Description of one condition (trait)

Names

Preferred name

The name of the condition used for reporting from ClinVar by default.

When available, this is a preferred term from SNOMED CT. Other sources may include Office of Rare Diseases Research (ORDR), Human Phenotype Ontology (HPO), OMIM®, and MeSH. The name for the condition that the submitter provides is retained, but is mapped to controlled vocabularies when possible. Because testing laboratories may know only the name of the ordered test (e.g. "deafness"), the indication for testing or the test name may be provided (see below), with condition submitted as "not provided". A list of disorder names used by ClinVar/GTR is provided from ClinVar's ftp site in the file disease_names.

When the condition is a drug response, the condition name should be constructed as the drug name + response.

Spreadsheet: Variant.Condition ID type, Condition ID value
Spreadsheet: Variant.Preferred condition name
Spreadsheet: Variant.Condition descriptions**XML:**
Trait/Name/ElementValueType/[@val_type = "name"] Preferred
XML: Trait/Name/ElementValue
db: GTR.clinvar.target_attr where attr_type = 17 (AttributeSet)

Content note: When an allele is asserted to be globally benign

When making a clinical assertion of "benign", the condition can either be listed as benign relative to a specific condition, or to the concept representing "all highly penetrant genetic disorders", indicating that the variant is asserted to be globally benign. For the latter, please submit the name of the condition as "AllHighlyPenetrant".

Alternate name(s)

- Optional, multiple allowed

Other names used for this condition. These are added to the set of search terms used in ClinVar, MedGen, and GTR.**Spreadsheet:** Variant.Condition description
Spreadsheet: Variant.Preferred condition name
XML: Trait/Name/ElementValueType/[@val_type = "name"] Alternate
XML: Trait/Name/ElementValue
db: GTR.clinvar.target_attr where attr_type = 18 (AttributeSet)

Preferred symbol

- Optional, only one allowed.

The preferred symbol for the condition. The final value used for display may be recalculated by NCBI.

Spreadsheet: not represented
XML: Trait/Symbol/ElementValueType/[@val_type = “name”] Preferred
XML: Trait/Symbol/ElementValue
db: GTR.clinvar.target_attr where attr_type = 19

Alternate symbols(s)

- Optional, multiple allowed.

Alternate symbols for the condition.

Spreadsheet: not represented
XML: Trait/Symbol/ElementValueType/[@val_type = “name”] Alternate
XML: Trait/Symbol/ElementValue
db: GTR.clinvar.target_attr where attr_type = 20

Indication for testing

- Optional.

Spreadsheet: Variant.Indication
Spreadsheet: CaseData.Indication
XML: //Sample/Indication
db: GTR.clinvar.version (as XML)

Attributes

These are based on the AttributeSet structure, and thus can be used to capture values assigned to defined information categories, along with supporting documentation. The values can be words, integers, decimals, and/or dates. Types are restricted by an enumerated list of allowed values per major information set. These restrictions may be applied in the XSD, or only in the underlying relational database. If you wish to suggest a new attribute, please contact us at clinvar@ncbi.nlm.nih.gov.

- Optional, multiple allowed

Concept	attr_type	XML	column in spreadsheet
Usual age of onset	257	MeasureTrait/AttributeSet/MeasureTraitAttributeType/[@val_type = “name”] AgeOfOnset	not represented

Reported penetrance	353	MeasureTrait/AttributeSet/MeasureTraitAttributeType/[@val_type = “name”] Penetrance	not represented
Mode of inheritance*	162	MeasureTrait/AttributeSet/MeasureTraitAttributeType/[@val_type = “name”] ModeOfInheritance	Variant.Mode of inheritance
Severity **	468,740	MeasureTrait/Severity MeasureTrait/ObservedIn/ObservedData/Severity	not represented
Activity level **	258	MeasureTrait/ObservedIn/ObservedData/[ObsAttributeType= “ActivityLevel”][Attribute]	not represented

* **NOTE**: mode of inheritance is stored as an attribute of the allele/condition relationship. If consistent for all alleles, mode of inheritance is also stored as an attribute of the condition itself.

****NOTE:** Severity and activity level, which may qualify the condition, can be submitted either at the level of the relationship between a set of conditions and a set of variations (*e.g*. for this variation, the condition is severe or the activity level of the product is decreased), or at the level of the observations themselves (*e.g*. for this sample, the condition is severe or the activity level of the product is decreased).

Category/Type of condition term

Category of the condition term. Only one classification is allowed per condition identifier. If the submission provides the name of a trait with a type different from ClinVar’s categorization, ClinVar retains what is submitted but continues to report the ClinVar categorization and reviews the discrepancy with the submitter.

The options and usage are tabulated at the beginning of the Condition section.

Spreadsheet: Variant.Condition category
XML: TraitSet/Trait/TraitType
db: GTR.clinvar.target.id_type

Indication

Testing labs may only know the indication for testing, not the full condition of the tested individual. Consistent with UMLS, ClinVar treats these as Findings.

Spreadsheet: CaseData.Indication
XML: Trait/Name/ElementValueType/[@val_type = “name”]
XML: ObservedIn/TraitSet/Trait/Name/ElementValue
db: GTR.clinvar.target_attr where attr_type = 17 (AttributeSet)

Variant allele(s)

ClinVar maintains information about sequence changes by representing location on an explicit reference sequence and the nucleotide or amino acid observed at that location. The allele may be the same as reference, for example polymorphic sites in which the reference sequence matches the allele about which

information is being submitted. Sequence changes may be single or multiple. Because it is possible to submit information about multiple sequence changes with relationship to condition, sequence variants are submitted as a set (MeasureSet), even if the size of the set is one.

Content notes

Names (XML)

Please submit official designations of alleles (e.g. CYP2D6*2) as Name of type="Preferred".

HGVS expressions (XML)

Please submit HGVS expressions as attributes of Type="HGVS". ClinVar extracts each expression submitted as an HGVS name, validates it (can that allele be identified on the referenced sequence), compute other HGVS expressions, and returns all values for public display. If a gene is known to have multiple splice variants, or legacy numbering systems, or more than one nucleotide change resulting in the same protein change, it is critical that the HGVS value contain the reference nucleotide sequence, the version of the reference sequence, the location, and the change. Submissions with insufficient specificity are returned for review or reported as non-validated.

Sets of variants

If a condition has been observed in a compound heterozygote, or is associated with a haplotype with more than one sequence change, the combination of alleles is represented as a set. This representation is to be distinguished from co-occurrence, which is used to report rare alleles in genes thought to contribute to a condition, but for which the alleles are not thought to be pathogenic in the reported context. Sets of variants can have many of the same attributes as single variants, *e.g.* names, identifiers in other databases, allele frequencies, *etc*. If HGVS is not used to represent whether the alleles occur on the same chromosome or not, then cis or trans must be supplied. ClinVar uses this information to indicate whether the set of variations defines a haplotype or a compound heterozygote.

Submitter's identifier for the allele being described in the submission

- optional

ClinVar uses this value, plus the condition, to construct a unique key for the clinical assertion being submitted. When a submission is processed successfully, ClinVar generates a report for the submitter based on the submitter's key, the condition description, and the accession assigned to the submission (SCV).

Spreadsheet: Variant.Local ID
XML: ClinvarSubmission/ClinvarSubmissionID/@localKey
XML: ClinvarSubmission/MeasureTrait/ExternalID/@id
db:GTR.clinvar.measure_target.local_key

URL to submitter's record

- optional

A URL that points to the submitted variant on the submitter's website.

Spreadsheet: Variant.URL

XML: Xrefs/@url (The Xref structure can be provided at many levels in the submission, to indicate what URL the submitter has for that object).
db: GTR.clinvar.attr_source

Definition of the variant by locations and sequence changes

- required

The reference sequence and version, such as NM_000492.3, NG_016465.3, NC_000007.13, LRG_76t1.

Spreadsheet: Variant.Reference sequence
XML: //ClinvarSubmission/MeasureSet/Measure/AttributeSet/MeasureAttributeType='HGVS'
XML: //ClinvarSubmission/MeasureSet/Measure/AttributeSet/Attribute
db: GTR.clinvar.mset + GTR.clinvar.msubset+GTR.clinvar.measure

The location and sequence change for the submitted variant.

Spreadsheet: Variant.HGVS
XML: //ClinvarSubmission/MeasureSet/Measure/AttributeSet/MeasureAttributeType='HGVS'
XML: //ClinvarSubmission/MeasureSet/Measure/AttributeSet/Attribute
db: GTR.clinvar.mset + GTR.clinvar.msubset+GTR.clinvar.measure

OMIM allelic variant ID

- optional

An OMIM allelic variant ID is reported for a set of variations as appropriate. If an allele occurs in more than one gene, and has multiple allelic variant ids, then both are reported. If curation determines that there are multiple allelic variant identifiers for the same allele, both identifiers are reported in that case as well.**Spreadsheet** Variant.Variation identifiers
XML: //ClinvarSubmission/MeasureSet/XRef
db: GTR.clinvar.mset_attr

Additional descriptors

Names and other attributes of a set of alleles can be submitted similar to the names and attributes of single alleles.

Strand

- optional

Representation of strand on which each allele of a set of alleles is found. Likely of concern only when explicitly representing a haplotype.**Spreadsheet:** not represented
XML: Measure.SequenceLocation@Strand
db: GTR.clinvar.seq_loc.strand

Each Variant allele

Each allele needs to be described unambiguously as the location of the variation and the sequence at that location. This requirement may be achieved in any of several ways.

Location

There are multiple options to specify the location of a variation. To permit unambiguous mapping to the genome, a submission in nucleotide coordinates, as accession.version+location is highly preferred. If a LRG sequence is used, the version is not applicable. If the description of the variation is provided via an HGVS expression which includes the explicit reference sequence and its version, then location need not be reported as a separate value. For chromosome locations, the assembly name and chromosome names (or accession+version) must be supplied for the sequence to be identifiable.

Cytogenetic

For variations defined by sequence, cytogenetic location is optional and can be provided by NCBI. For large structural variations defined only cytogenetically, this is required.

Spreadsheet: Variant.Chromosome
XML: Measure/CytogeneticLocation
db: GTR.clinvar.seq_loc.cytogenetic + GTR.clinvar.seq_loc.chr

Nucleotide Location

The location of this variant defined by chromosome and position. The location may be a point or a range, with or without defined end points. If a point, only start needs to be provided, but ClinVar computes stop based on the value reported as start. For variants without exact locations defined, multiple values are provided to represent the boundaries of what is known (e.g. outer start and outer stop, inner start and inner stop). These are defined as documented here:
http://www.ncbi.nlm.nih.gov/dbvar/content/overview/

Spreadsheet: Variant/Start/ReferenceAllele/AlternateAllele, Outer start, Outer stop, etc.
XML: Measure/SequenceLocation with multiple attributes to define the assembly, sequence, and position/boundaries of the variation's location
db: GTR.clinvar.seq_loc (multiple columns)

Nucleotide change as HGVS

The nucleotide change for a variation represented as an HGVS expression. For more details about how NCBI and ClinVar manage HGVS expressions, please see
http://www.ncbi.nlm.nih.gov/clinvar/docs/hgvs_types.

The Reference sequence must include reference and version, such as NM_000492.3, NG_016465.3, NC_000007.13. The Variation name is the c., g., m., n. or r. portion of the full HGVS expression and must be in agreement with the reference type.

Spreadsheet: Variant.Reference sequence and Variant.HGVS
XML: Measure/AttributeSet/Attribute/MeasureAttributeType = "HGVS"
XML: Measure/AttributeSet/Attribute/Attribute
db: GTR.clinvar.measure_attr

Protein change (HGVS, single letter or 3 letter amino acid abbreviations)

- optional

Although submitting the definition of a variation *only* in protein coordinates is accepted, this format is not recommended. It is our goal to map sequence variation to the genome, and protein coordinates are not always sufficient. That said, submission of a variation as both the nucleotide change and protein change is desirable, to support confirmation of location.**Spreadsheet:** Variant.Alternate designations
XML: Measure/AttributeSet/Attribute/@Type="HGVS"
db: GTR.clinvar.measure_attr

Official variant name

- optional

This must be an official allele name. For variants that are assigned official allele names, e.g. CYP3A4*18 for one of the cytochrome P450 gene CYP3A4; or HLA-DRA*0102 for the MHC gene HLA-DRA.**Spreadsheet:** Variant.Official allele name
XML: Measure/Name@type=preferred
db: GTR.clinvar.measure_attr.attr_char where attr_type = 17 (AttributeSet)

Alternate names

- optional

Other names in common use for an allele. **Spreadsheet:** Variant.Alternate designations
XML: Measure/Name/ElementValueType=alternate
XML: Measure/Name/ElementValue
db: GTR.clinvar.measure_attr.attr_char where attr_type = 18 (AttributeSet)

Identifiers in public databases

- optional

  Identifiers in dbSNP/dbVar/OMIM, locus-specific databases, *etc.* Special handling is provided for identifiers generated by NCBI, namely rs#, nsv, nssv, in that they have dedicated attribute types and are stored in measure_attr. Other non-NCBI public identifiers are stored in attr_source. At

times, a submission may include information that the location of a variation can be identified by an rs# or an nsv# or some other public identifier.

Spreadsheet: Variant. Variation identifiers
XML: Measure.AttributeSet.Attribute.rsNumber
XML: Measure.AttributeSet.Attribute.nsv
Measure/XRef/@DB
db: GTR.clinvar.attr_source
or
db: GTR.clinvar.measure_attr

Location relative to a gene, protein, or other genomic location

- optional

Some of these values are based on sequence ontology terms and computed per transcript. Content can be computed by NCBI and/or provided by submitter. This category includes exon and intron numbers, position relative to splicing or regulatory regions, position in conserved protein domains, *etc*. The sequence ontology terms used by NCBI include:

- UTR (SO:0000203)
    - 5_prime_UTR (SO:0000204)
    - 3_prime_UTR (SO:0000205)
- Upstream location
    - Upstream variant (SO:0001631)
        - Within 5kb (SO:0001635)
        - Within 2kb (SO:0001636)
- Downstream location
    - downstream_gene_variant (SO:0001632)
        - 5KB_downstream_variant (SO:0001633)
        - 500B_downstream_variant (SO:0001634)
- Splice site
    - splice_site (SO:0000162)
- Distance from nearer splice junction
  (can be calculated if not provided)
- Regulatory site (yes/no or name of promoter/locus control region)
    - Promoter: SO:0000167

Intron or exon number

- optional

Submitters may provide an intron or exon designation and Arabic numeral (*e.g*., exon 4, intron 3, not IVS 3 or Exon IV). The sequence used to define the numbering system for the intron/exon organization should also be included.**Spreadsheet:** Variant.Location
XML: Measure/AttributeSet/Attribute Type='Location'
db: GTR.clinvar.measure_attr where attr_type = 472

Region name (active site, conserved domain, unspecified, etc)

- optional

Submitters may provide a domain name in which the variation is found. NCBI will also report when the variation lies within a known domain.

Spreadsheet: Not represented**XML:** Measure/AttributeSet/Attribute@type='Domain'
db: GTR.clinvar.mset_attr where attr_type = 473

Total exons in transcript

This optional concept is included in the dictionary because the value may be included in our public displays. The data are not to be submitted however, and will be provided based on the sequence used to define any gene annotation.

Other regions with similar sequence which may confound interpretation

Submitters may describe other regions in the genome with sequence highly similar to the context of the reported variant, and which may affect variation calls. This attribute may describe a gene or a variant.

Spreadsheet: Not represented
XML: MeasureSet/AttributeSet/Attribute@type='RelatedSequence'
db: GTR.clinvar.mset_attr

Molecular consequence

- optional

Molecular consequence is reported from sequence ontology terms when available, and, when possible, are computed per transcript by NCBI. These terms are in this group because they can be calculated explicitly from the type and location of the variation, unlike the functional consequence which must be established experimentally (or predicted). (AttributeSet).

Spreadsheet: Not represented
XML: Measure/AttributeSet/Attribute type='MolecularConsequence'
db. GTR.clinvar.measure_attr

Comment about molecular consequence

As with most other data elements, a free text comment may be submitted about the molecular consequence. The comment structure should be used if the consequence being submitted is not defined by the Sequence Ontology group. We strongly recommend, however, that an SO term be requested if current terms are insufficient.

Spreadsheet: Variant.Comment
XML: Measure/AttributeSet/Attribute Type='MolecularConsequence'
Measure/AttributeSet/Comment/CommentText

Functional consequence

- optional

These attributes are provided by the submitter since they require determination of the consequences of the molecular change. Each is qualified by whether the submitter predicted the consequence or established it experimentally. Options include terms used by the LOVD databases and terms established by VariO (variationontology.org) and Sequence Ontology (SO, http://www.sequenceontology.org/browser/obo.cgi).

- Loss of function
- Gain of function
- Overexpression
- Underexpression
- Splice_site_lost
- Splice_site_gained
- Nonsense mediated decay
- affects function
- probably affects function
- probably does not affect function
- does not affect function
- unknown

Spreadsheet: Variant.Functional consequence
XML: Measure/AttributeSet/Attribute@Type='FunctionalConsequence'
db: GTR.clinvar.measure_attr where attr_type = 474

Method for determining functional consequence

Spreadsheet: FunctionalEvidence. Method
XML: Method/Type
db: GTR.clinvar.method.method_type

Functional consequence comment

- optional

As with most other data elements, a free text comment may be submitted about the functional consequence. The comment structure should be used if the consequence being submitted is not defined by VariO, SO, or LOVD. We strongly recommend, however, that a term be requested if current terms are insufficient.

Spreadsheet: Variant.Comment on functional consequence

XML: Measure/AttributeSet/Attribute Type='FunctionalConsequence'
MeasureAttributeSet/Comment/CommentText
db: GTR.clinvar.measure_attr

Type of variation

Description of the type of variation, using terms from the Sequence Ontology as appropriate. Note that the option *undefined* exists as a default value. NCBI reassigns the type when necessary. For the list of allowed values, please refer to the xsd or the authorities document (http://www.ncbi.nlm.nih.gov/clinvar/docs/authorities/).

Spreadsheet: Variant.Variant type (required for structural variants)
XML: Measure/MeasureType
db: GTR.clinvar.measure.id_type

Description of the asserted relationship between a set of conditions and a set of variations

Mode of inheritance

- optional

The mode of inheritance is reported as an attribute of the relationship between a variation (or a set of variations) and the disorder. The ontology being used is Human Phenotype Ontology (HPO). The list of allowed values is maintained for ClinVar and GTR on GTR's ftp site:

ftp://ftp.ncbi.nlm.nih.gov/pub/GTR/standard_terms/Mode_of_inheritance.txt

If you provide other, please specify, e.g. other:'new mode of inheritance'. **Spreadsheet:** Variant.Mode of inheritance
XML: MeasureTrait/AttributeSet/MeasureTraitAttributeType = "ModeOfInheritance"
XML: MeasureTrait/AttributeSet/Attribute
db: GTR.clinvar.mt_attr where attr_type = 163

Clinical significance

Clinical significance is also reported explicitly for co-occurring variations which may contribute to the condition. See the co-occurrence section.

Note: When an allele is asserted to be globally benign
If a submitted allele appears to be benign for all highly penetrant disorders (rather than a specific

disorder), ClinVar is representing the condition by the name 'AllHighlyPenetrant'. See the Condition section.

Description

- optional

Clinical significance is required for public reporting. One option, however, is 'not provided' so that submitters are not required to calculate significance to submit their data. The list of allowed values is maintained for ClinVar and GTR on GTR's ftp site:

ftp://ftp.ncbi.nlm.nih.gov/pub/GTR/standard_terms/Clinical_significance.txt

If you provide are going to provide "other" or if you have a value that you think should be included in this list please contact us at clinvar@ncbi.nlm.nih.gov. **Spreadsheet**: Variant.Clinical significance
XML: MeasureTrait/ClinicalSignificance/Description
db: GTR.clinvar.mt_attr where attr_type = 151

Date last evaluated

- Required if available

Date the clinical significance was last evaluated.
Spreadsheet: Variant.Date last evaluated
XML: MeasureTrait/ClinicalSignificanceDateLastEvaluated
db: GTR.clinvar.mt_attr.attr_date where attr_type = 151

Assertion method

- required

The method that was used to make the assertion of clinical significance. Terms for assertion method are being developed; for now, we are using the same value as provided for Collection method.

Spreadsheet: Variant.Assertion method
XML: MeasureTrait/Assertion/Method
db:

Citations

- optional

Citations documenting the assertion of clinical significance. Any of PubMed, PubMedCentral, DOI, NCBI Bookshelf combined with the id in that database (e.g. PMID:123456). See Citations.
Spreadsheet: Variant.Clinical significance citations
Spreadsheet: Variant. Citations or URLs for clinical significance without database identifiers
XML: MeasureTrait/ClinicalSignificance/Citation (CitationType)
db: GTR.dbo.citation; GTR.dbo.citation_many

Comment

- Optional, but highly recommended

Free text describing the assertion of clinical significance. See Comments.**Spreadsheet**:
Variant.Comment on clinical significance

XML: MeasureTrait/ClinicalSignificance/Comment

db GTR.dbo.comment

Custom Assertion Score

- optional

Submitter-specific scoring method names and the values obtained for each, (where submitter has alternate system/nomenclature for clinical significance). These are not standardized, not stored in a normalized fashion in relational columns, but are being retained for the submitter's use.**Spreadsheet**:
Not represented in spreadsheet

XML: MeasureTrait.CustomAssertionScore[@Value= "*string*"]

XML: MeasureTrait/CustomAssertionScore/CustomAssertionScoreType

db: GTR.clinvar.version.xml_object

Citations

See Citations.

Spreadsheet: Not represented in spreadsheet

XML: MeasureTrait/CustomAssertionScore/Citation

XRef

See XRefs.

Spreadsheet: Not represented in spreadsheet

XML: MeasureTrait/CustomAssertionScore/XRef

Evidence/ObservedIn

The evidence section maintains the details necessary to review the medical importance of a set of variations with respect to a diagnosis or medical outcome. This evidence may be computational, based on experimental testing, or observations in human subjects. A submission may contain multiple observations for the same allele/condition combination. In the XML, these are represented by multiple //MeasureTrait/ObservedIn elements; in the spreadsheet these are represented by multiple lines in the CaseData tab with the same value in the Linking ID column as the LinkingID column of the Variant tab.

Spreadsheet: Variant, CaseData

XML: ObservedIn

db: GTR.clinvar.observations

Sample

This section is used to describe what was studied to generate the submission and its supporting evidence.

Species

- required

Defaults to human if not supplied.

Spreadsheet: FunctionalEvidence.Species
XML: ObservedIn/Sample.[Species[@TaxonomyId=9606] = “human”]
db: clinvar.sample.txid

Cell line

- optional (required if evidence was generated in a cell line)

Name of the cell line. To be used only when supporting experimental evidence is being reported.**Spreadsheet**: FunctionalEvidence.Cell line
XML: ObservedIn/Sample/CellLine
db: clinvar.sample.cell_line

Strain/breed

- optional

Name of the strain or breed that was analyzed in this experimental study**Spreadsheet**: FunctionalEvidence.Strain/breed
XML: ObservedIn/Sample/Strain
db: clinvar.sample.strain

Allele origin

- required

Spreadsheet: Variant.Allele origin
Spreadsheet: CaseData.Allele origin
XML: ObservedIn/Sample/Origin
db: clinvar.sample.origin

Age range

- optional

The range of ages included in this sample. If age range is an important variable in your submission, with different observations based on the age, please submit each observation separately, rather than lumping into one summary observation with one sample description**Spreadsheet**: Variant.Age range

Spreadsheet: CaseData.Age
XML: ObservedIn/Sample/Age
db: clinvar.sample.min_age
db: clinvar.sample.max_age
db: clinvar.sample.age_units

Geographic origin

- optional

Can be used to indicate country, continent, or a region in which this allele was reported.**Spreadsheet**: Variant.Geographic origin

Spreadsheet: CaseData.Geographic origin
XML: ObservedIn/Sample/GeographicOrigin
db: GTR.clinvar.sample.geographic_origin

Population Group/**Ethnicity**

- optional

Name or description of the ethnicities of the individual in which the allele was reported.**Spreadsheet**: Variant.Ethnicity

Spreadsheet: CaseData.Ethnicity
XML: ObservedIn/Sample/Ethnicity
db: clinvar.sample.ethnicity

Tissue

- optional

Name or description of the tissue that was assayed. Highly recommended if the origin is somatic or if an experimental analysis. **Spreadsheet**: Variant.Tissue

Spreadsheet: CaseData.Tissue
XML: ObservedIn/Sample/Tissue
db: clinvar.sample.tissue

Fraction of sample which is tumor-containing

- optional

Free text description of the fraction of the sample that has tumor cells. Applicable only if origin is somatic.

Spreadsheet: Not represented in spreadsheet
XML: ObservedIn/Sample/FractionTumor

db: clinvar.sample.fraction_tumor

Affected status

- required

Indicates whether the individual or sample in which the variant was identified had the condition for which an assertion is being made. Accepted values: yes, no, unknown. The values "yes" and "no" should be used when the sample is restricted to individuals older than the expected age of onset. Otherwise submit 'unknown'. **Spreadsheet:** Variant.Affected status

Spreadsheet: CaseData.Affected status
XML: ObservedIn/Sample/AffectedStatus
db: clinvar.sample.affected_status

Number of chromosomes tested

- optional, but highly recommended

Spreadsheet: Not represented in spreadsheet
XML: ObservedIn/Sample/NumberChrTested
db: clinvar.sample.chr_tested

Number of individuals tested

- optional

The number of subjects on which this submission is based.**Spreadsheet**: Variant.Total number of individuals tested
XML: ObservedIn/Sample/NumberTested
db: clinvar.sample.individuals_tested

Sex

- optional

If explicit numbers are known in a sample set, they should be specified in NumberMales and /or NumberFemales. Otherwise, use Sex.**Spreadsheet**: Variant.Sex
Spreadsheet: CaseData.Sex
XML: ObservedIn/Sample/NumberMales
XML: ObservedIn/Sample/NumberFemales
XML: ObservedIn/Sample/Gender
db: clinvar.sample.males
db: clinvar.sample.females
db. clinvar.obs_attr

Number families tested

- optional

Spreadsheet: Variant.Number of families tested

XML: ObservedIn/FamilyData/FamilyHistory[@NumFamilies= "*integer*"]
db: clinvar.sample.families_tested

Number of families with segregation observed

- optional

Spreadsheet: Variant. Number of families with Segregation observed
XML: ObservedIn/FamilyData/FamilyHistory[@NumFamiliesWithSegregation= "*integer*"]
db: clinvar.obs_attr

Family history

- optional

Used to indicate that at least one other member of a family has the reported condition. It does not require that other family members were included in the observation set. **Spreadsheet**: Variant.Family history
XML: ObservedIn/Sample/FamilyData/FamilyHistory
db: clinvar.sample.positive_family_history

Number of Independent Affected Subjects tested

- Optional but highly desirable

ClinVar computes how many times a variant has been seen in affected individuals from independent families by data aggregated by the submitter (Variant tab on the spreadsheet), or by summing data submitted as cases .**Spreadsheet:** CaseData.Proband, Family ID (CaseData)
Spreadsheet: Number of families with variant, Affected status=yes
XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "IndependentObservations"
XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
XML: ObservedIn.Sample.AffectedStatus = "yes"
db: GTR.clinvar.obs_attr where attr_type =IndependentObservations
db: GTR.clinvar.sample.affected_status='yes'

Family ID

- optional

Spreadsheet: CaseData.Family ID
XML: ObservedIn/Sample/FamilyInfo@PedigreeID
db: clinvar.sample.family_id

Citations

- optional

Citation documenting the evidence. Any of PubMed, PubMedCentral, DOI, NCBI Bookshelf combined with the id in that database (e.g. PMID:123456). See Citations.**Spreadsheet**: Variant.Evidence citations
Spreadsheet: CaseData.Evidence citations
XML: ObservedIn/ObservedData/Citation
db: GTR.dbo.citation and GTR.dbo.citation_many

Method

Test name or type

- optional

Because testing laboratories may know only the name of the ordered test (e.g. "deafness"), the test name may be submitted with condition submitted as "not provided". A GTR id maybe submitted instead of a free text name.

Spreadsheet: Variant.Test name or type
Spreadsheet: CaseData.Test name or type
XML: ObservedIn/Method/Citation/XRef (for GTR ids)
db: clinvar.method.extrn_src (for GTR ids)
db: clinvar.method.extrn_id (for GTR ids)

Observations/Observed Data

ClinVar uses 'observations' to store evidence generated from a combination of methods applied to a sample. Observations are also used to represent some conclusions about results from a combination of methods, such as 'confirmed by independent methods'.

XML path = /ClinvarSubmissonSet/ClinvarSubmisson/MeasureTrait/ObservedIn

Methods for data collection

The method used to gather data for a submission is reported as a method type. The options are used to support evaluation of the submission, as well as to allow representation of clear distinctions between primary data and data culled from the literature.

Spreadsheet: Variant.Collection method
Spreadsheet: CaseData.Collection method
XML: MeasureTrait/ObservedIn/Method
db:

Option	Explanation
Literature only	Data extracted from published literature with interpretation as reported in the citation
Clinical testing	Data provided from genetic testing, interpretation as currently provided by the tester. Interpretation may be guided from the literature, but the number of individuals tested are reported only from the direct testing
Reference population	Data gathered from baseline studies of a population group of apparently unaffected individuals to assess allele frequencies
Case-control	Data gathered from a research setting to compare alleles observed in cases and controls (without data about segregation)
Research	General research method when other more specific methods to not apply
In vivo	
In vitro	
Not provided	

Number of families with variant

- optional

Spreadsheet: Variant.Number of families with variant
XML: ObservedIn/FamilyData/FamilyHistory[@NumFamiliesWithVariant= "*integer*"]
db: clinvar.obs_attr

Number of Individuals having a variant allele

- optional

Spreadsheet: Variant.Number of individuals with variant
XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "VariantAlleles"
XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
OR
XML: ObservedIn/ObservedData/Attribute/
db: GTR.clinvar.obs_attr where attr_type ='VariantAlleles'

Number of independent individuals with variant

- optional

Spreadsheet: not represented
XML:
db: clinvar.obs_attr

Number of chromosomes with variant

- optional

Spreadsheet: Variant.Number of chromosomes with variant
XML:
db: clinvar.obs_attr

Number of individuals who have only the variant [homozygous and hemizygous]

- optional

Spreadsheet: Variant.Number of homozygotes
XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SubjectsOnlyVariant"
XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
db: GTR.clinvar.obs_attr where attr_type ='SubjectsOnlyVariant'

Number of single heterozygotes

- optional

This count includes single heterozygotes reported in the context of dominant mode of inheritance, and single heterozygotes observed (in a recessive context) but where no other pathogenic variant was identified to classify as a compound heterozygote.**Spreadsheet**: Variant.Number of single heterozygotes
XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SingleHeterozygote"
XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*
db: GTR.clinvar.obs_attr where attr_type ='SingleHeterozygote'

Number of compound heterozygotes

- optional

This is a count of heterozygotes where another heterozygous pathogenic variant partner WAS identified. Both variant alleles must be submitted.**Spreadsheet**: Variant.Number of compound heterozygotes

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "CompoundHeterozygote"

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

db: GTR.clinvar.obs_attr where attr_type =CompoundHeterozygote

Number of individuals with *de novo* variant observed

- optional

Spreadsheet: Variant.Number of individuals with de novo variant observed

XML:

db: clinvar.obs_attr

Mosaicism

- optional

Spreadsheet: Variant.Mosaicism

Spreadsheet: CaseData.Mosaicism

XML:

db: clinvar.obs_attr

Number of affected subjects with genotype consistent with mode of inheritance

- optional

The sum of single heterozygotes, compound heterozygotes, and homozygotes for the reported allele with a condition consistent with the asserted mode of inheritance**Spreadsheet**: not represented

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "GenotypeAndMOIConsistent"

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type =GenotypeAndMOIConsistent

Number of affected subjects with this variant who also have another variant thought to be responsible for condition

- optional

This information is captured to evaluate pathogenicity, based on the logic that if another allele may account for the observed condition, this one has unknown pathogenicity.**Spreadsheet**: not represented

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "SubjectsWithDifferentCausativeAllele"

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type=' SubjectsWithDifferentCausativeAllele'

Number of instances observed of heterozygous parent transmitting normal allele to an affected child

- optional

One of several observations that support evaluation of penetrance.**Spreadsheet**: not represented

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "HetParentTransmitNormalAllele "

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type ='HetParentTransmitNormalAllele '

db: GTR.clinvar.sample.affected_status='yes'

Number of independent families demonstrating co-segregation

- optional

Spreadsheet: Not represented

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "CosegregatingFamilies"

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

XML: ObservedIn.Sample.AffectedStatus = "yes"

db: GTR.clinvar.obs_attr where attr_type ='CosegregatingFamilies'

db: GTR.clinvar.sample.affected_status='yes'

Number of informative meioses

- optional

Spreadsheet: Not represented

XML: ObservedIn/ObservedData/ObsAttributeType[@val_type="name"] = "InformativeMeioses"

XML: ObservedIn/ObservedData/Attribute/@integerValue = *number*

db: GTR.clinvar.obs_attr where attr_type ='InformativeMeioses'

Co-occurrence

- optional

Submitters may have identified other alleles which may contribute to the condition and were observed in the individuals studied, but which are not included in the interpretation. The values provided in each co-occurrence set describe the distinct sets of co-occurring alleles and the number of times each was observed.

- Zygosity (optional)
- (*Value*=HomozygousVariant, SingleHeterozygote, or CompoundHeterozygote)

XML: ObservedIn/Co-occurrenceSet/Zygosity
db: GTR.clinvar.co_occurs.zygosity_type

- Count of individuals with this co-occurrence set

XML: ObservedIn/Co-occurrenceSet/Count
db: GTR.clinvar.co_occurs.num_reported

- Description of the alleles observed (co-occurring genotypes)

Spreadsheet: CaseData.Co-occurrences, same gene
Spreadsheet: CaseData.Co-occurrences, other genes
XML: ObservedIn/Co-occurrenceSet/AlleleDescrSet/Name
db: GTR.clinvar.alelle_set.name

- Orientation

XML: ObservedIn/Co-occurrenceSet/AlleleDescrSet/RelativeOrientation
db: GTR.clinvar.alelle_set.orientation

- Zygosity

XML: ObservedIn/Co-occurrenceSet/AlleleDescrSet/Zygosity
db: GTR.clinvar.alelle_set.zygosity_type

- Clinical Significance

XML: ObservedIn/Co-occurrenceSet/AlleleDescrSet/ClinicalSignificance
db: GTR.clinvar.alelle_set.clinical_significance

Description of a gene

- optional

A gene, if provided, must be unambiguously defined. That definition may be supplied either by the official symbol (see the Names section), or an identifier in a public database: GeneID, HGNC id or MIM number (see the Attributes section). The gene is considered optional because the location of the variation should be sufficient to define the gene.

Genes are represented in ClinVar in two major contexts:

1. The gene in which a variation has been described
2. An explicit representation of gene-condition relationship.

The former case is represented in the XML as a MeasureRelationship of type 'variant in gene'.

The latter case is used internally, but is not currently processed as a publicly reported accession.

Names

ClinVar coordinates the representation of names of genes with NCBI's Gene database. In other words, the name is defined primarily by the nomenclature established by the HUGO Gene Nomenclature Committee (HGNC).

Preferred name

optional, only one allowedThe preferred full name as reported by NCBI's Gene database.

Spreadsheet: not represented
XML: MeasureRelationship/Name/ElementValueType[@val_type="name"] = Preferred
XML: MeasureRelationship/Name/ElementValue
db: GTR.clinvar.measure_attr where attr_type = 17 (AttributeSet)

Alternate name(s)

- optional, multiple allowed

Other names used for this gene, as provided by NCBI's Gene database.

Spreadsheet: not represented
XML: MeasureRelationship/Name/ElementValueType[@val_type="name"] = Alternate
XML: MeasureRelationship/Name/ElementValue
db: GTR.clinvar.measure_attr where attr_type = 18 (AttributeSet)

Preferred symbol

- optional, only one allowed

The official symbol from HGNC. This may be used by submitters to indicate the gene.

Spreadsheet: Variant.Gene symbol

XML: MeasureRelationship/Symbol/ElementValueType[@val_type="name"] = Preferred
XML: MeasureRelationship/Symbol/ElementValue
db: GTR.clinvar.measure_attr where attr_type = 19 (AttributeSet)

Alternate symbols(s)

- optional, multiple allowed

Alternate gene symbols from Gene, as provided by NCBI's Gene database.

Spreadsheet: not represented
XML: MeasureRelationship/Symbol/ElementValueType[@val_type="name"] = Alternate
XML: MeasureRelationship/Symbol/ElementValue
db: GTR.clinvar.measure_attr where attr_type = 20 (AttributeSet)

Attributes

- Examples are GeneID, HGNC id, MIM number, chromosome, cytogenetic band, chromosome sequence location, and whether or not there are pseudogenes or paralogs. The set of optional attributes is designed to capture information necessary to set the framework for interpretation of variation.
- Many of these attributes are not duplicated in the ClinVar database but are provided by NCBI as imports from the Gene database or defined by the sequence used to define the gene structure.
- GeneID, HGNC id, and MIM number may be used by submitters to indicate the gene.

Overview of Gene related concepts reported by ClinVar

Concept	attr_type	XML	column in spreadsheet
GeneID		XRef/@DB=Gene	Not represented
HGNC id		XRef/@DB=HGNC	Not represented
MIM number		XRef/@DB=OMIM	Not represented
Location of the gene on the GRC	Location	Measure/SequenceLocation/@Assembly, @Chr, @start, @stop	SubmissionInfo.Assembly; Variant.Chromosome coordinates

assembly; chromosome			
Gene relationships		Measure/MeasureRelationship/AttributeSet/@Type=”gene relationship”	

Has paralogs

Optional flag that variation calls in this region may be confounded by paralogs in the genome. In other words, this field is not intended to report locations of all paralogs for this location; but a warning, projected to be computed by NCBI, that paralogs exist.

Spreadsheet: Not represented
XML: Measure.MeasureRelationship.AttributeSet.Attribute/@Type=”ParalogInfo”
db: GTR.clinvar.measure_attr

Has pseudogenes

Optional flag which can be provided by the submitter, but usually computed by NCBI, that a gene has pseudogenes. In other words, this field is not intended to report locations of all pseudogenes for a gene; but a warning that pseudogenes exist which may affect confidence in variation calls in this region.

Spreadsheet: Not represented
XML: Measure.MeasureRelationship.AttributeSet.Attribute/@Type=”PseudoGeneInfo”
db: GTR.clinvar.measure_attr