

Package: RClinVarbitration (via r-universe)

July 21, 2026

Type Package

Title Stream ClinVar XML into Relational DuckDB Tables

Version 0.1.1

Description Streams official 'ClinVar' VCV XML releases, including gzip inputs, into focused relational tables in 'DuckDB'. A package-owned native extension uses libxml2's forward reader to retain VCV variants, normalized alleles, RCV aggregates, SCV submissions, conditions, observations, citations, and attributable evidence text without loading the release into R. ClinVarbitration policy is a derived SQL layer over those retained source relations, with a separate direct reproducer for archived NCBI flat-file releases.

License GPL (>= 2)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 8.0.0

Imports DBI

Suggests duckdb, knitr, rmarkdown, roxygen2, tinytest

VignetteBuilder knitr

SystemRequirements libxml2 (>= 2.9.0), zlib, pkg-config, C compiler;
Rtools (Windows)

URL <https://github.com/RGenomicsETL/RClinVarbitration>,
<https://r genomicsetl.github.io/RClinVarbitration/>

BugReports <https://github.com/RGenomicsETL/RClinVarbitration/issues>

Config/pak/sysreqs libxml2-dev pkg-config zlib1g-dev

Repository <https://r genomicsetl.r-universe.dev>

Date/Publication 2026-07-21 20:31:49 UTC

RemoteUrl <https://github.com/RGenomicsETL/RClinVarbitration>

RemoteRef HEAD

RemoteSha 1700ec9a53c3d9c350a24ced8c4a673207633382

Contents

rclinvarbitration_download_clinvar	2
rclinvarbitration_enable	3
rclinvarbitration_export_clinvarbitration_parquet	4
rclinvarbitration_extension_path	5
rclinvarbitration_import_xml	5
rclinvarbitration_init	6
rclinvarbitration_policy_sql	7
rclinvarbitration_policy_version	7
rclinvarbitration_reproduce_clinvarbitration_parquet	8
rclinvarbitration_schema_sql	9
Index	10

rclinvarbitration_download_clinvar
<i>Download official ClinVar source files</i>

Description

Downloads either the current ClinVar release or a named monthly archive from the NCBI HTTPS service. The VCV XML file is the normal input to [rclinvarbitration_import_xml\(\)](#). The two flat files are optional, validation-only inputs to [rclinvarbitration_reproduce_clinvarbitration_parquet\(\)](#).

Usage

```
rclinvarbitration_download_clinvar(  
  release = "latest",  
  file = "vcv_xml",  
  cache_dir = tools::R_user_dir("RclinVarbitration", "cache"),  
  overwrite = FALSE,  
  quiet = FALSE  
)
```

Arguments

release	"latest" or a monthly archive in "YYYY-MM" form.
file	One or more of "vcv_xml", "submission_summary", and "variant_summary".
cache_dir	Directory in which downloaded files are cached.
overwrite	Download even when a valid cached file is present?
quiet	Passed to utils::download.file() .

Details

Existing files form a local cache. Current files and archived VCV XML files are checked against NCBI's MD5 sidecar; a matching file is reused and a stale or corrupt file is replaced only after a temporary download is verified. NCBI does not publish MD5 sidecars beside archived flat files, so those files are reused by name unless `overwrite = TRUE`.

A complete VCV release is several gigabytes. The default cache location is returned by `tools::R_user_dir("RclinVarbitration", "cache")`. The function raises R's download timeout to at least the value of option `RclinVarbitration.download_timeout` (7200 seconds by default) for the call.

Value

A named character vector of local paths. The download attribute is a data frame containing source URLs, MD5 values when available, and cache hit status.

Examples

```
xml <- rclinvarbitration_download_clinvar(release = "latest")
xml_2026_03 <- rclinvarbitration_download_clinvar(release = "2026-03")
```

rclinvarbitration_enable

Enable native ClinVar XML scanning on a DuckDB connection

Description

Loads the package-owned rclinvarbitration extension. Its native `clinvar_xml_entities(path)` table function and `rclinvar_json_field()` scalar are the compact, ClinVar-specific one-pass staging surface used by `rclinvarbitration_import_xml()`. The connection must have been created with `duckdb::duckdb(config = list(allow_unsigned_extensions = "true"))`, as for any locally built DuckDB extension.

Usage

```
rclinvarbitration_enable(con, extension_path = NULL)
```

Arguments

`con` A DuckDB DBI connection.

`extension_path` Optional explicit exact-version extension path.

Value

`con`, invisibly.

rclinvarbitration_export_clinvarbitration_parquet

Export an allele-level ClinVarbitration-compatible Parquet file

Description

Writes the seven columns in Centre for Population Genomics ClinVarbitration's `clinvar_decisions.tsv`: `contig`, `position`, `reference`, `alternate`, `clinical_significance`, `gold_stars`, and `allele_id`. The source is the package's allele-level policy view, not the disease-level decision view, so each output row is usable as a locus/alleles annotation record. Both GRCh37 and GRCh38 are supported. The output retains every qualifying source locus, including distinct X/Y locations for one AlleleID.

Usage

```
rclinvarbitration_export_clinvarbitration_parquet(
  con,
  path,
  release_id,
  assembly = c("GRCh38", "GRCh37"),
  profile_id = "default",
  submitter_exclusions = character()
)
```

Arguments

<code>con</code>	A DuckDB DBI connection initialized with <code>rclinvarbitration_init()</code> .
<code>path</code>	New output <code>.parquet</code> file path.
<code>release_id</code>	Imported ClinVar release label to export.
<code>assembly</code>	Genome assembly: "GRCh38" or "GRCh37".
<code>profile_id</code>	Policy profile identifier, normally "default".
<code>submitter_exclusions</code>	Additional submitter names to exclude from this export. Matching is case-insensitive and ignores surrounding whitespace. These exclusions are combined with any exclusions already stored for <code>profile_id</code> ; imported source submissions are not deleted.

Details

The file is schema-compatible with the upstream TSV/Hail decision resource, but is not claimed to be byte-for-byte equivalent: this package derives submissions and locations from VCV XML, whereas upstream uses ClinVar's tab-delimited submission and variant summaries. PM5 is deliberately not exported; Rduckhts/DuckHTS own downstream consequence and PM5 processing.

Value

A named list describing the written Parquet file, invisibly.

rclinvarbitration_extension_path

Locate a version-matched RClinVarbitration DuckDB extension

Description

The package uses DuckDB's unstable C extension ABI for its streaming table function, so it bundles one artifact per supported exact engine release.

Usage

```
rclinvarbitration_extension_path(duckdb_version = NULL, duckdb_platform = NULL)
```

Arguments

duckdb_version Exact DuckDB version, with or without a v prefix. NULL selects the version reported by the installed duckdb package.

duckdb_platform

Exact platform reported by PRAGMA platform. NULL queries a temporary connection to the installed duckdb package. DuckDB distinguishes windows_amd64 from R-devel's windows_amd64_mingw even though both artifacts contain the same MinGW-built machine code.

Value

An absolute extension path.

rclinvarbitration_import_xml

Stream a ClinVar VCV XML release into relational DuckDB tables

Description

The native extension reads .xml and .xml.gz with a libxml2 forward reader. One compact JSON-backed row per selected ClinVar entity is written to a disk-backed staging table in one XML pass, projected into the public ClinVar relations without an EAV pivot, and dropped. No XML DOM, XML blob, generic parser-node graph, or R data-frame materialization is used. Each public relation is committed independently, the release catalogue row marks completion, and failed imports remove partial rows for release_id.

Usage

```
rclinvarbitration_import_xml(
  con,
  path,
  release_id,
  replace = FALSE,
  source_url = NULL,
  source_md5 = NULL
)
```

Arguments

con	A DuckDB DBI connection.
path	Path to an official ClinVar VCV XML or XML.GZ release.
release_id	User-supplied release label stored with every row.
replace	Replace rows already stored for release_id?
source_url	Optional source URL for the release catalogue. When path is returned directly by rclinvarbitration_download_clinvar() , its download metadata supplies this value automatically.
source_md5	Optional 32-character source MD5 digest. Download metadata is used automatically when available.

Value

A named numeric vector with imported entity counts.

```
rclinvarbitration_init
```

Initialize the ClinVar relational schema

Description

Initialize the ClinVar relational schema

Usage

```
rclinvarbitration_init(con)
```

Arguments

con	A DuckDB DBI connection.
-----	--------------------------

Value

con, invisibly.

`rclinvarbitration_policy_sql`*ClinVarbitration policy SQL*

Description

Returns DuckDB views implementing the supported policy over clinvar_disease_submissions. The policy bins submitted classifications, excludes unknown bins and applies the upstream policy's intended qualified Illumina benign exclusion, optionally applies profile-specific submitter exclusions, prefers evidence evaluated from 2016 onward while always retaining expert-panel and practice guideline evidence, and applies the 60/20 majority rule. The disease-level and allele-level views both reproduce the upstream strong-review rule: the first retained practice-guideline or expert-panel classification in source order is decisive.

Usage

```
rclinvarbitration_policy_sql(  
  policy_version = rclinvarbitration_policy_version()  
)
```

Arguments

`policy_version` Exact supported policy version. See [rclinvarbitration_policy_version\(\)](#).

Details

`clinvar_policy_pathogenic_alleles` is the disease-specific pathogenic or likely-pathogenic join surface for downstream Rduckhts/DuckHTS annotation. It does not perform VEP or PM5 computation.

Value

A named character vector of DuckDB CREATE OR REPLACE VIEW statements.

`rclinvarbitration_policy_version`*Current ClinVarbitration policy version*

Description

The identifier pins the policy semantics adapted from Centre for Population Genomics ClinVarbitration 2.2.11 (upstream commit 658b9f241eb2d43aa11214b153b19c1e18a16337) and records that decisions are grouped per disease rather than only per variation. The identifier has no package-local suffix: the view names, rather than a speculative v1, state whether the output is disease-scoped or allele-scoped.

Usage

```
rclinvarbitration_policy_version()
```

Value

A character scalar.

```
rclinvarbitration_reproduce_clinvarbitration_parquet
```

Reproduce ClinVarbitration decisions from archived ClinVar flat files

Description

Reproduces Centre for Population Genomics ClinVarbitration 2.2.11's allele-level decision algorithm directly from NCBI's versioned submission_summary and variant_summary files. The generated Parquet has the exact seven-column clinvar_decisions.tsv layout: contig, position, reference, alternate, clinical_significance, gold_stars, and allele_id.

Usage

```
rclinvarbitration_reproduce_clinvarbitration_parquet(
  con,
  submission_path,
  variant_path,
  path,
  assembly = c("GRCh38", "GRCh37"),
  submitter_exclusions = character()
)
```

Arguments

con	A DuckDB DBI connection.
submission_path	Archived NCBI submission_summary_YYYY-MM.txt.gz.
variant_path	Archived NCBI variant_summary_YYYY-MM.txt.gz.
path	New output .parquet path.
assembly	Genome assembly: "GRCh38" or "GRCh37".
submitter_exclusions	Submitter names to exclude for a blinded run.

Details

This is separate from [rclinvarbitration_import_xml\(\)](#) and the disease-aware policy views. It exists to reproduce and differentially validate the upstream allele-level artifact on matching archived flat-file releases. It uses DuckDB's streaming CSV reader and SQL aggregation; it does not invoke Python, Hail, VEP, or PM5 logic.

Value

A named list describing the written Parquet file, invisibly.

`rclinvarbitration_schema_sql`

ClinVar relational schema SQL

Description

Returns DuckDB DDL for the focused ClinVar schema. Public identifiers and domain relations are retained directly: VCV variants, alleles and assembly locations, genes, RCV aggregates, SCV submissions, conditions, observations, citations, attributes, and attributable discovery text. XML parser nodes are not persisted. The release catalogue enforces its small primary key; the release-scale analytical tables expose logical key columns without DuckDB ART indexes so complete imports remain memory-bounded.

Usage

```
rclinvarbitration_schema_sql()
```

Value

A named character vector of SQL statements.

Index

`rclinvarbitration_download_clinvar`, [2](#)
`rclinvarbitration_download_clinvar()`,
 [6](#)
`rclinvarbitration_enable`, [3](#)
`rclinvarbitration_export_clinvarbitration_parquet`,
 [4](#)
`rclinvarbitration_extension_path`, [5](#)
`rclinvarbitration_import_xml`, [5](#)
`rclinvarbitration_import_xml()`, [2](#), [3](#), [8](#)
`rclinvarbitration_init`, [6](#)
`rclinvarbitration_init()`, [4](#)
`rclinvarbitration_policy_sql`, [7](#)
`rclinvarbitration_policy_version`, [7](#)
`rclinvarbitration_policy_version()`, [7](#)
`rclinvarbitration_reproduce_clinvarbitration_parquet`,
 [8](#)
`rclinvarbitration_reproduce_clinvarbitration_parquet()`,
 [2](#)
`rclinvarbitration_schema_sql`, [9](#)

`utils::download.file()`, [2](#)