CTIS Database Documentation

Overview

Database Type: SQLite
File Name: `ctis.db`
Purpose: Store clinical trial data scraped from the European Medicines Agency's Clinical Trial Information System (CTIS)
Schema Version: 1.3.0
Last Updated: November 2024

The CTIS database is a normalized relational database that stores comprehensive clinical trial information including trial metadata, eligibility criteria, endpoints, products, sites, contacts, and regulatory status across EU member states.

Database Configuration

SQLite Pragmas

```sql
PRAGMA journal_mode=WAL;        -- Write-Ahead Logging for better concurrency
PRAGMA busy_timeout=30000;      -- 30-second timeout for locked database
PRAGMA foreign_keys=ON;         -- Enforce foreign key constraints
```

Indexing Strategy

All tables include strategic indexes for:

- Primary keys (automatic)
- Foreign keys (ctNumber references)
- Frequently queried fields (countries, dates, status)
- Common filter columns (medical conditions, phases, trial status)

Database Schema

Table Relationships

```
trials (master table)
  ├── inclusion_criteria (1:many)
  ├── exclusion_criteria (1:many)
  ├── endpoints (1:many)
  ├── trial_products (1:many)
  ├── trial_sites (1:many)
  ├── trial_people (1:many)
```

```
├── ms_status (1:many)
├── country_planning (1:many)
├── site_contact (1:many)
├── trial_funding (1:many)
├── trial_scientific_advice (1:many)
└── trial_relationships (1:many)
```

Primary Key: All tables use `ctNumber` (European Clinical Trial Number) as the foreign key to link to the master `trials` table.

Table 1: `trials`

Purpose: Master table containing core trial information and metadata.

Schema

Column	Type	Description
ctNumber	TEXT PRIMARY KEY	European Clinical Trial Number (unique identifier)
ctStatus	INTEGER	Internal status code
ctPublicStatusCode	TEXT	Public-facing status code
title	TEXT	Full official trial title
shortTitle	TEXT	Short title / Protocol code
sponsor	TEXT	Sponsoring organization name
trialPhase	TEXT	Trial phase (Phase I, II, III, IV, etc.)
therapeuticAreas	TEXT	Comma-separated MeSH therapeutic area codes
medicalCondition	TEXT	Primary medical condition being studied
medicalConditionsList	TEXT	Complete list of conditions
isConditionRareDisease	INTEGER	Boolean: 1 if rare disease, 0 otherwise
conditionMeddraCode	TEXT	MedDRA medical condition code
conditionMeddraLabel	TEXT	MedDRA label (human-readable)
conditionSynonyms	TEXT	Alternative condition names
conditionAbbreviations	TEXT	Common abbreviations
countries	TEXT	Comma-separated list of countries
decisionDate	TEXT	Authorization decision date (YYYY-MM-DD)
publishDate	TEXT	Publication date on CTIS (YYYY-MM-DD)
lastUpdated	TEXT	Last update timestamp (YYYY-MM-DD)

Feasibility Fields

Column	Type	Description
ageCategories	TEXT	Target age ranges (comma-separated)
isPediatric	INTEGER	Boolean: includes pediatric population
isAdult	INTEGER	Boolean: includes adult population
gender	TEXT	Gender eligibility (Male/Female/All)
isRandomised	INTEGER	Boolean: randomized trial
blindingType	TEXT	Blinding method (Open/Single/Double/Triple)
trialScope	TEXT	Geographic scope
mainObjective	TEXT	Primary trial objective
primaryEndpointsCount	INTEGER	Number of primary endpoints
secondaryEndpointsCount	INTEGER	Number of secondary endpoints
estimatedRecruitmentStartDate	TEXT	Planned recruitment start (YYYY-MM-DD)
estimatedEndDate	TEXT	Planned trial end date (YYYY-MM-DD)

Disclosure Timing Fields

Column	Type	Description
trialCategory	TEXT	Trial category (1, 2, or 3)
expectedDosageDisclosureDate	TEXT	When dosage info becomes public
dosageVisibleNow	INTEGER	Boolean: dosage currently visible

Enhanced Fields (v1.3.0)

Column	Type	Description
who_utn	TEXT	WHO Universal Trial Number
nct_number	TEXT	ClinicalTrials.gov NCT identifier
isrctn_number	TEXT	ISRCTN registry number
additional_registry_ids	TEXT	Other registry identifiers (JSON)
pip_number	TEXT	Pediatric Investigation Plan number
pip_decision_date	TEXT	PIP decision date
is_transition_trial	INTEGER	Boolean: EudraCT→CTIS transition
eudract_number	TEXT	Previous EudraCT number
global_end_date	TEXT	Global trial end date
allocation_method	TEXT	Randomization method
number_of_arms	INTEGER	Number of treatment arms

System Fields

Column	Type	Description
data_json	TEXT	Complete raw JSON from CTIS API
updated_at_utc	TEXT	UTC timestamp of last database update

Indexes

- `idx_trials_lastUpdated` on `lastUpdated`
- `idx_trials_medicalCondition` on `medicalCondition`
- `idx_trials_phase` on `trialPhase`
- `idx_trials_nct` on `nct_number`
- `idx_trials_pip` on `pip_number`
- `idx_trials_eudract` on `eudract_number`

Key Constraints

- `ctNumber` is PRIMARY KEY (unique, not null)
- Foreign keys are enforced when `PRAGMA foreign_keys=ON`

Table 2: `inclusion_criteria`

Purpose: Eligibility criteria that participants must meet to join the trial.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
criterionNumber	INTEGER	Sequential number of criterion
criterionText	TEXT NOT NULL	Full text of inclusion criterion

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- `idx_inclusion_ct` on `ctNumber`

Usage Notes

- Each row represents one inclusion criterion
- Typical trials have 5-20 inclusion criteria
- Order is preserved via `criterionNumber`

Table 3: `exclusion_criteria`

Purpose: Conditions or factors that prevent individuals from participating.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
criterionNumber	INTEGER	Sequential number of criterion
criterionText	TEXT NOT NULL	Full text of exclusion criterion

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- `idx_exclusion_ct` on `ctNumber`

Table 4: `endpoints`

Purpose: Primary and secondary outcome measures.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
endpointType	TEXT NOT NULL	"primary" or "secondary"
endpointNumber	INTEGER	Sequential number within type
endpointText	TEXT NOT NULL	Description of what is measured
timeFrame	TEXT	When endpoint is assessed

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- `idx_endpoints_ct` on `ctNumber`
- `idx_endpoints_type` on `endpointType`

Usage Notes

- Primary endpoints: main outcomes (typically 1-3)
- Secondary endpoints: supporting outcomes (typically 5-15)

Table 5: `trial_products`

Purpose: Investigational medicinal products used in the trial.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
productRole	TEXT	Role (Test product/Reference/Placebo)
productName	TEXT	Commercial or generic name
activeSubstance	TEXT	Active pharmaceutical ingredient
atcCode	TEXT	Anatomical Therapeutic Chemical code
pharmaceuticalForm	TEXT	Dosage form (Tablet/Injection/etc)
route	TEXT	Route of administration
maxDailyDose	TEXT	Maximum daily dose (numeric)
maxDailyDoseUnit	TEXT	Dose unit (mg/IU/mL)
maxTreatmentPeriod	INTEGER	Maximum treatment duration
maxTreatmentPeriodUnit	TEXT	Period unit (Day/Week/Month)
isPaediatric	INTEGER	Boolean: pediatric indication
isOrphanDrug	INTEGER	Boolean: orphan drug designation
authorizationStatus	TEXT	Marketing authorization status
raw_json	TEXT	Complete raw product JSON

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- `idx_products_ct` on `ctNumber`

Table 6: `trial_sites`

Purpose: Geographic locations where the trial is conducted.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
site_name	TEXT	Name of the clinical trial site
organisation	TEXT	Operating organization name
country	TEXT	Country name

Column	Type	Description
city	TEXT	City name
address	TEXT	Street address
postal_code	TEXT	Postal/ZIP code
path	TEXT	Internal path reference
raw_json	TEXT	Complete raw site JSON

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, site_name, organisation, country, city, address, postal_code, path)

Indexes

- idx_sites_ct on ctNumber
- idx_sites_country on country

Table 7: trial_people

Purpose: Contact information for trial personnel.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
name	TEXT	Full name of contact person
role	TEXT	Role (PI/Coordinator/Public Contact)
email	TEXT	Email address
phone	TEXT	Phone number
country	TEXT	Country location
city	TEXT	City location
site_name	TEXT	Associated trial site
organisation	TEXT	Organization affiliation
path	TEXT	Internal path reference
raw_json	TEXT	Complete raw contact JSON

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, name, role, email, phone, country, city, site_name, organisation, path)

Indexes

- idx_people_ct on ctNumber

Table 8: ms_status

Purpose: Member State-specific trial status and timeline information.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
member_state	TEXT NOT NULL	EU Member State code
status	TEXT	Current status in this country
decision_date	TEXT	Authorization decision date
start_date	TEXT	Trial start date
recruitment_start	TEXT	Recruitment start date
recruitment_end	TEXT	Recruitment end date
temporary_halt	TEXT	Temporary halt date (if any)
restart_date	TEXT	Restart date (after halt)
end_date	TEXT	Trial end date
early_termination_date	TEXT	Early termination date (if applicable)
early_termination_reason	TEXT	Reason for early termination
last_update	TEXT	Last status update date
captured_at	TEXT NOT NULL	When this data was captured
row_hash	TEXT	Hash for change detection

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, member_state, status, captured_at)

Indexes

- idx_ms_status_ct on ctNumber
- idx_ms_status_country on member_state
- idx_ms_status_country_status on (member_state, status)
- idx_ms_status_recruiting on (status, recruitment_start)

Usage Notes

- Tracks country-specific trial status

- Allows historical tracking via `captured_at`
- Critical for understanding trial recruitment by geography

Table 9: `country_planning`

Purpose: Planned participant enrollment by country.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
country	TEXT NOT NULL	Country name
planned_participants	INTEGER	Number of planned participants

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, country)

Indexes

- `idx_country_planning_ct` on `ctNumber`

Table 10: `site_contact`

Purpose: Detailed site contact information including principal investigators.

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
country	TEXT	Country code or name
org_name	TEXT	Organization name
site_name	TEXT	Trial site name
address	TEXT	Full address
city	TEXT	City
postal_code	TEXT	Postal code
pi_name	TEXT	Principal Investigator name
pi_email	TEXT	PI email address
pi_phone	TEXT	PI phone number

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, country, org_name, site_name, pi_name, pi_email)

Indexes

- idx_site_contact_ct on ctNumber

Table 11: trial_funding

Purpose: Funding source information (Enhanced v1.3.0).

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
funding_source_type	TEXT	Type of funding source
funding_source_name	TEXT	Name of funding organization
funding_source_country	TEXT	Country of funding source
is_primary_funder	INTEGER	Boolean: primary funder flag

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- idx_funding_ct on ctNumber

Table 12: trial_scientific_advice

Purpose: Scientific advice received from regulatory authorities (Enhanced v1.3.0).

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
advice_authority	TEXT	Regulatory authority providing advice
advice_type	TEXT	Type of advice received
advice_date	TEXT	Date advice was received

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)

Indexes

- `idx_advice_ct` on `ctNumber`

Table 13: `trial_relationships`

Purpose: Relationships between trials (Enhanced v1.3.0).

Schema

Column	Type	Description
id	INTEGER PRIMARY KEY AUTOINCREMENT	Unique record identifier
ctNumber	TEXT NOT NULL	Foreign key to trials table
related_ctNumber	TEXT NOT NULL	Related trial number
relationship_type	TEXT	Type of relationship (e.g., "Extension", "Sub-study")
description	TEXT	Description of relationship

Constraints

- FOREIGN KEY (ctNumber) REFERENCES trials(ctNumber)
- UNIQUE constraint on (ctNumber, related_ctNumber, relationship_type)

Indexes

- `idx_relationships_ct` on `ctNumber`
- `idx_relationships_related` on `related_ctNumber`

Data Types and Conventions

Date Format

- **Standard:** ISO 8601 format `YYYY-MM-DD`
- **Timestamps:** ISO 8601 with time `YYYY-MM-DDTHH:MM:SSZ`
- **Empty dates:** NULL or empty string

Boolean Fields

- **Storage:** INTEGER (0 = false, 1 = true)
- **Common fields:** isPediatric, isAdult, isConditionRareDisease, isRandomised

List Fields

- **Format:** Comma-separated values
- **Examples:**

- `countries`: "France, Germany, Spain"
- `therapeuticAreas`: "C04, C14, C20"

JSON Fields

- **Purpose:** Store complete raw API responses
- **Encoding:** UTF-8
- **Fields:** data_json, raw_json

Common Queries

Get All Trial Information

```sql
SELECT * FROM trials WHERE ctNumber = 'CT-EU-00012345';
```

Get Active Trials in a Country

```sql
SELECT DISTINCT t.*
FROM trials t
JOIN ms_status ms ON t.ctNumber = ms.ctNumber
WHERE ms.member_state = 'Spain'
  AND ms.status = 'Authorised'
  AND ms.recruitment_start IS NOT NULL;
```

Get Trials by Medical Condition

```sql
SELECT ctNumber, title, sponsor, medicalCondition
FROM trials
WHERE medicalCondition LIKE '%diabetes%'
  OR conditionMeddraLabel LIKE '%diabetes%';
```

Get Trials with Eligibility Criteria

```sql
```

```sql
SELECT
    t.ctNumber,
    t.title,
    t.ageCategories,
    t.gender,
    COUNT(DISTINCT ic.id) AS inclusion_count,
    COUNT(DISTINCT ec.id) AS exclusion_count
FROM trials t
LEFT JOIN inclusion_criteria ic ON t.ctNumber = ic.ctNumber
LEFT JOIN exclusion_criteria ec ON t.ctNumber = ec.ctNumber
GROUP BY t.ctNumber
HAVING inclusion_count > 0;
```

Get Trial Sites by Country

```sql
SELECT
    ts.ctNumber,
    t.title,
    ts.country,
    ts.city,
    ts.site_name,
    ts.organisation
FROM trial_sites ts
JOIN trials t ON ts.ctNumber = t.ctNumber
WHERE ts.country = 'France'
ORDER BY ts.city, ts.site_name;
```

Get Recruitment Status Summary

```sql
SELECT
    ms.member_state,
    ms.status,
    COUNT(*) AS trial_count
FROM ms_status ms
GROUP BY ms.member_state, ms.status
ORDER BY ms.member_state, trial_count DESC;
```

Data Integrity

Foreign Key Enforcement

```sql
```

```
PRAGMA foreign_keys=ON;
```

This ensures:

- Cannot delete trial without deleting related records
- Cannot insert criteria/endpoints/etc without valid ctNumber

Unique Constraints

Most child tables have UNIQUE constraints to prevent duplicate records:

- Sites: unique by (ctNumber, site_name, organisation, country, city, address)
- People: unique by (ctNumber, name, role, email, phone, country, city)
- MS Status: unique by (ctNumber, member_state, status, captured_at)

Transaction Safety

All database operations use transactions:

```python
conn.execute("BEGIN TRANSACTION")
# ... operations ...
conn.commit()
```

Performance Optimization

Indexing Strategy

1. **Primary Keys:** Automatic indexing
2. **Foreign Keys:** Explicit indexes on all ctNumber columns
3. **Filter Columns:** Indexes on frequently filtered fields (country, status, phase)
4. **Composite Indexes:** Multi-column indexes for common query patterns

WAL Mode

Write-Ahead Logging (WAL) provides:

- Better concurrency (readers don't block writers)
- Improved performance for read-heavy workloads
- More reliable database in case of crashes

Query Optimization Tips

1. Always use `WHERE ctNumber = ?` for single-trial queries (uses PRIMARY KEY)
2. Use `JOIN` instead of multiple queries when possible

3. Add indexes for custom query patterns
4. Use `EXPLAIN QUERY PLAN` to analyze slow queries

Database Size Estimates

Trials	Database Size	Notes
100	~5-10 MB	Small test dataset
1,000	~50-100 MB	Typical research project
10,000	~500 MB - 1 GB	Full CTIS extract
50,000+	~2-5 GB	Complete historical archive

Size varies based on:

- Number of criteria per trial (5-50 typical)
- Number of endpoints (2-20 typical)
- Number of sites (1-200+ for multicenter)
- JSON data retention (raw_json fields)

Maintenance Operations

Vacuum Database

```sql
VACUUM;
```

Rebuilds database file to reclaim space and improve performance.

Analyze Statistics

```sql
ANALYZE;
```

Updates query optimizer statistics for better query plans.

Check Integrity

```sql
PRAGMA integrity_check;
```

Verifies database file integrity.

Version History

v1.3.0 (November 2024)

- Added `trial_funding` table
- Added `trial_scientific_advice` table
- Added `trial_relationships` table
- Enhanced trials table with WHO UTN, NCT number, ISRCTN
- Added PIP tracking fields
- Added transition trial fields

v1.2.0

- Added `ms_status` tracking with historical captures
- Added `country_planning` table
- Added `site_contact` table with PI details

v1.1.0

- Enhanced trials table with feasibility fields
- Added disclosure timing fields
- Improved indexing strategy

v1.0.0

- Initial schema with core tables
- Basic trial information and criteria

Migration to Aurora

Recommended Aurora Schema

For Aurora (PostgreSQL-compatible), consider these modifications:

1. **Change INTEGER PRIMARY KEY to SERIAL:**

```sql
id SERIAL PRIMARY KEY
```

2. **Add UUID columns for global uniqueness:**

```sql
uuid UUID DEFAULT gen_random_uuid()
```

3. **Use JSONB instead of TEXT for JSON fields:**

```sql
data_json JSONB
raw_json JSONB
```

4. **Add timestamps:**

```sql
created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
updated_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
```

5. **Create views for common queries**
6. **Implement partitioning for ms_status by captured_at**
7. **Add materialized views for dashboard queries**

Related Files

- **Python Module:** `ctis_database.py` - Database operations
- **Configuration:** `ctis_config.py` - Database path and settings
- **Processor:** `ctis_processor.py` - Data insertion logic
- **Excel Export:** `ctis_excel_gen.py` - Excel generation from database
- **Metadata:** `CTIS_Key_Mappings.json` - Code-to-label mappings

Support and References

- **CTIS Portal:** https://euclinicaltrials.eu/
- **EMA Information:** https://www.ema.europa.eu/en/human-regulatory/research-development/clinical-trials-information-system
- **Source Code:** https://hendrik.codes/post/scraping-the-clinical-trials-information-system

---

This documentation describes the SQLite database schema as of November 2024. Schema may evolve as CTIS requirements change.