IME 312 Clinical Research Database

Part 1A Deliverable

Cyrus Mak, Myron Chen, Noah Pidding

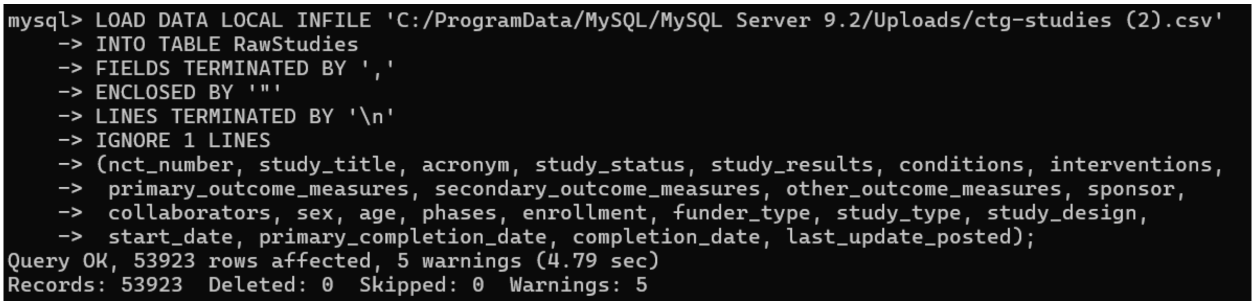
Table of Contents

Project Scope

ICU Today seeks to provide a database that keeps up to date research and clinical trials for doctors, and researchers for ease of use. With clinical trials being updated constantly, the ability to provide better patient care is improved every day. We aim to create a database system to keep doctors and patients in the loop about publically funded government trials.

https://clinicaltrials.gov/

Establish entries, attributes, and relations



```
mysql> LOAD DATA LOCAL INFILE 'C:/ProgramData/MySQL/MySQL Server 9.2/Uploads/ctg-studies (2).csv'
    -> INTO TABLE RawStudies
    -> FIELDS TERMINATED BY ','
    -> ENCLOSED BY '"'
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 LINES
    -> (nct_number, study_title, acronym, study_status, study_results, conditions, interventions,
    ->  primary_outcome_measures, secondary_outcome_measures, other_outcome_measures, sponsor,
    ->  collaborators, sex, age, phases, enrollment, funder_type, study_type, study_design,
    ->  start_date, primary_completion_date, completion_date, last_update_posted);
Query OK, 53923 rows affected, 5 warnings (4.79 sec)
Records: 53923  Deleted: 0  Skipped: 0  Warnings: 5
```

CREATE DATABASE clinical_trials;

USE clinical_trials;

```
DROP TABLE RawStudies;
CREATE TABLE RawStudies (
    nct_number INT,             – Note to remove “NCT” str from whole column
    study_title TEXT,
    acronym VARCHAR(100),
    study_status VARCHAR(100),
    study_results VARCHAR(10),
    conditions TEXT,
    interventions TEXT,
    primary_outcome_measures TEXT,
    secondary_outcome_measures TEXT,
    other_outcome_measures TEXT,
    sponsor VARCHAR(255),
    collaborators TEXT,
    sex VARCHAR(20),
    age VARCHAR(50),
    phases VARCHAR(50),
    enrollment VARCHAR(50),
    funder_type VARCHAR(100),
    study_type VARCHAR(100),
    study_design TEXT,
    start_date DATE,
    primary_completion_date DATE,
    completion_date DATE,
    last_update_posted DATE,
);

SELECT * FROM clinical_trials.RawStudies;
mysql -u root -p --local-infile=1
```

[CODE WIP SUBJECT TO CHANGE]

Business Rules

1. One and only one "STUDY" will have one and only one "STUDY_DETAILS".
2. One "STUDY" can have one or many "SPONSORS".
3. "STUDY_DEMOGRAPHIC_MAP" can contain one or many "PARTICIPANTS".
4. "STUDY" can have one or many "STUDY_DEMOGRAPHIC_MAPS".
5. Each "STUDY" can have one and only one "SET OF MEASURES".
6. "MEASURES" as super type.
7. Other, primary, and secondary measures as sub-types of "MEASURES".
8. Every instance of “MEASURES” must have at least one “PRIMARY OUTCOME MEASURES”.
9. "MEASURES" can have one or no set of "SECONDARY OUTCOME MEASURES".

10. "MEASURES" can have one or no set of "OTHER OUTCOMES MEASURES".
11. An instance of the supertype measure can belong to more than one subtype (Overlapping).
12. A measure might not be classified.
13. "STUDY_ID" as "NCT_NUMBER".

Modeling Notes

1. **1:1 Relationship** between **STUDY** and **STUDY_DETAIL** via study_id.
2. **1:M Relationship** between **STUDY** and **SPONSOR**
3. **1:M Relationship** between **STUDY** and **STUDY_DEMOGRAPHIC_MAP**
4. **M:M Relationship** between **STUDY** and **PARTICIPANT** via **STUDY_DEMOGRAPHIC_MAP**
5. **1:1 Relationship** between **STUDY** and **MEASURE**
6. **Supertype/Subtype inheritance** with overlapping and optional categorization between **MEASURE**, **PRIMARY, SECONDARY,** and **OTHER**

Entity Definitions

STUDY: Represents a registered clinical trial. This is the core entity linking to all associated details like demographics, sponsors, and outcome measures.

- study_id (PK): Unique identifier of the study (e.g., NCT number).
- study_title: Official name or title of the study.
- study_result: Indicates whether results have been posted.
- study_conditions: The medical conditions or diseases being studied.
- measures_id (FK): Links to the set of measures associated with this study.
- sponser_id (FK): Links to the ***SPONSOR****(s)* of the study.

STUDY_DETAIL: Captures detailed metadata and operational characteristics for one and only one study.

- study_details_id (PK): Unique identifier for the detail record.
- study_details_study_status: The current status of the study (e.g., Recruiting, Completed).
- study_details_start_date: Date the study began.
- study_details_p_comp_date: Date of primary outcome data collection completion.
- study_details_comp_date: Date of final study completion.
- study_details_last_update: Last posted update date for the study.
- study_details_study_phase: The clinical trial phase (PHASE I, II, III, IV or NA).
- study_details_study_type: Type of study (e.g., Interventional, Observational).
- study_details_study_design: Description of study methodology.
- study_details_enrollment: Estimated or actual number of participants.
- study_details_interventions: Medical or procedural interventions applied.
- study_id (FK): Foreign key to the associated ***STUDY***.

SPONSOR: An organization or individual responsible for funding and managing a study.

- sponser_id (PK): Unique sponsor ID.
- sponsor_name: Name of the main sponsor(s)
- sponsor_collaborators: List or description of collaborating organization(s).
- sponsor_funder_type: Classification of funder (e.g., NIH, Industry, Other).

STUDY_DEMOGRAPIC_MAP: Junction table mapping studies to their participant profiles. Supports many-to-many relationships.

- study_id (PK, FK): Foreign key to ***STUDY.***
- profile_id (PK, FK): Foreign key to ***PARTICIPANT.***

PARTICIPANT: Demographic profile of individuals eligible for or involved in a study.

- profile_id (PK): Unique identifier for the participant demographic profile.
- partcipants_sex: Biological sex of participants (e.g., Male, Female, All).
- participants_age_group: Age range category of participants (e.g., Child, Adult, Older Adult).

MEASURE (Supertype): Generalized outcome data tracked across studies. Can represent different kinds of outcomes and can overlap (i.e., a measure may be primary and secondary).

- measures_id (PK): Unique ID for the measure set.
- primary_outcome_id (FK): Link to the ***PRIMARY_OUTCOME_MEASURE (required).***
- secondary_outcome_id (FK): Link to the ***SECONDARY_OUTCOME_MEASURE (optional*****).**
- other_outcome_id (FK): Link to the ***OTHER_OUTCOMES_MEASURE (optional).***

PRIMARY_OUTCOME_MEASURE (Subtype): Critical measure used to assess the primary objectives of the study.

- primary_outcome_id (PK): Unique ID for the primary outcomes
- secondary_outcome_details: Description of the secondary outcome

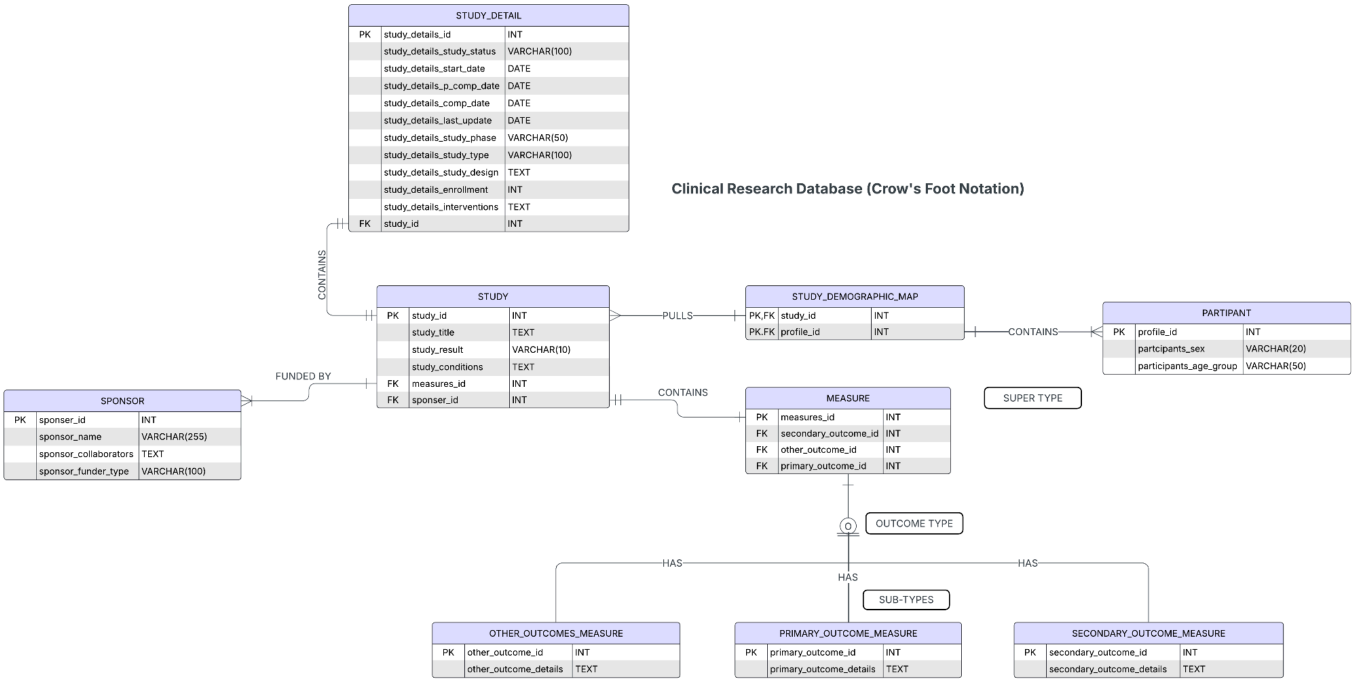
SECONDARY_OUTCOME_MEASURE (Subtype): Supplementary metrics evaluated in the study, not the main focus.

- secondary_outcome_id (PK): Unique ID for the secondary outcome.
- secondary_outcome_details: Description of the secondary outcome.

OTHER_OUTCOMES_MEASURE (Subtype): Additional or exploratory outcomes beyond the primary and secondary objectives.

- other_outcome_id (PK): Unique ID for the other outcome.
- other_outcome_details: Description of the other outcome..

Pre-Normalized E-R Diagram



Clinical Research Database (Crow's Foot Notation)

STUDY_DETAIL		
PK	study_details_id	INT
	study_details_study_status	VARCHAR(100)
	study_details_start_date	DATE
	study_details_p_comp_date	DATE
	study_details_comp_date	DATE
	study_details_last_update	DATE
	study_details_study_phase	VARCHAR(50)
	study_details_study_type	VARCHAR(100)
	study_details_study_design	TEXT
	study_details_enrollment	INT
	study_details_interventions	TEXT
FK	study_id	INT

STUDY		
PK	study_id	INT
	study_title	TEXT
	study_result	VARCHAR(10)
	study_conditions	TEXT
FK	measures_id	INT
FK	sponser_id	INT

STUDY_DEMOGRAPHIC_MAP		
PK,FK	study_id	INT
PK.FK	profile_id	INT

PARTIPANT		
PK	profile_id	INT
	partcipants_sex	VARCHAR(20)
	participants_age_group	VARCHAR(50)

SPONSOR		
PK	sponser_id	INT
	sponsor_name	VARCHAR(255)
	sponsor_collaborators	TEXT
	sponsor_funder_type	VARCHAR(100)

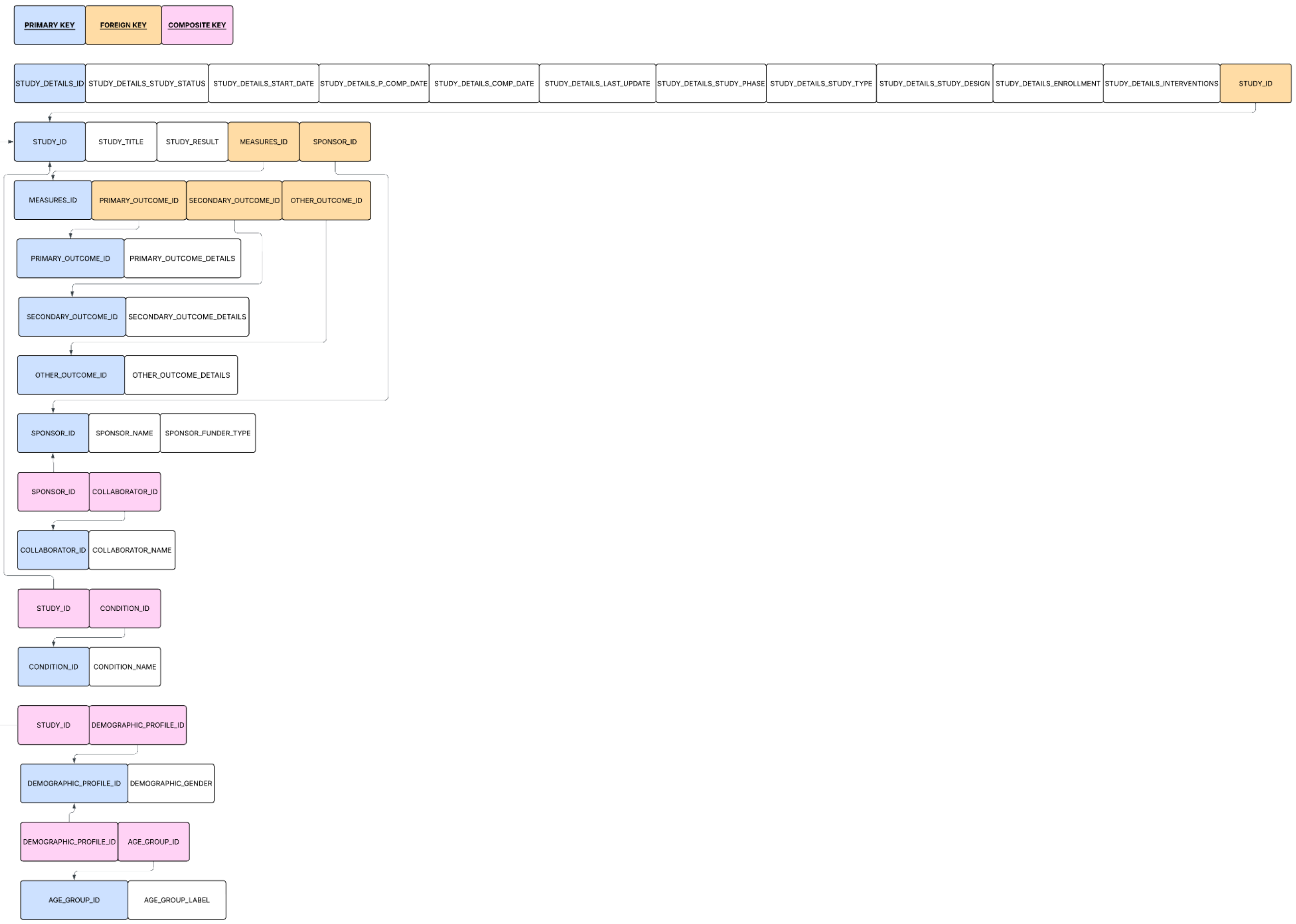
MEASURE		
PK	measures_id	INT
FK	secondary_outcome_id	INT
FK	other_outcome_id	INT
FK	primary_outcome_id	INT

OTHER_OUTCOMES_MEASURE		
PK	other_outcome_id	INT
	other_outcome_details	TEXT

PRIMARY_OUTCOME_MEASURE		
PK	primary_outcome_id	INT
	primary_outcome_details	TEXT

SECONDARY_OUTCOME_MEASURE		
PK	secondary_outcome_id	INT
	secondary_outcome_details	TEXT

Normalization



Previously, our rows from sponsor_collaborators, study_conditions, and demographic_age_group contained multiple values separated by "|" or commas. The table below provides an example of rows pre-normalization.

SPONSOR_COLLABORATORS	STUDY_CONDITION	DEMOGRAPHIC_AGE_GROUP
Northwestern University|National Institute on Aging (NIA)|University of Washington	Adenocarcinoma|Pancreatic Neoplasms|Neoplasm, Glandular|Neoplasms|Neoplasms Pancreatic|Digestive System Neoplasm|Endocrine Gland Neoplasms|Digestive System Disease|Pancreatic Diseases|Endocrine System Diseases	ADULT, OLDER ADULT, CHILD

These fields were multivalued and non-atomic, violating **1NF.** This also meant that they weren't normalized for lookup, querying or reusability, violating **3NF.**

To resolve these issues, separate junction tables were created for each of the 3 categories to record the relationships of the multiple values. **Note that all text in parentheses within tables are there for reference and will not actually be included in the tables.**

AGE_GROUP Table (New) and DEMOGRAPHIC Table are linked together through a many to many relationship through the DEMOGRAPHIC_AGE_GROUP Junction table (New)

DEMOGRAPHIC_PROFILE_ID	AGE_GROUP_ID
1001	1 (CHILD)
1001	2 (ADULT)
1002	3 (OLDER_ADULT)

- This ensured participants can be easily queried by age group
- AGE_GROUP values can be consistently reused across the database
- Data integrity (No typos like "Adult", "ADULT", "adult")

STUDY Table and CONDITION Table (New) are linked together through a many-to-many relationship through the STUDY_CONDITION (New) Junction Table.

STUDY_ID	CONDITION_ID
NCT001	10 (Asthma)
NCT001	12 (Allergies)
NCT002	11 (Diabetes)

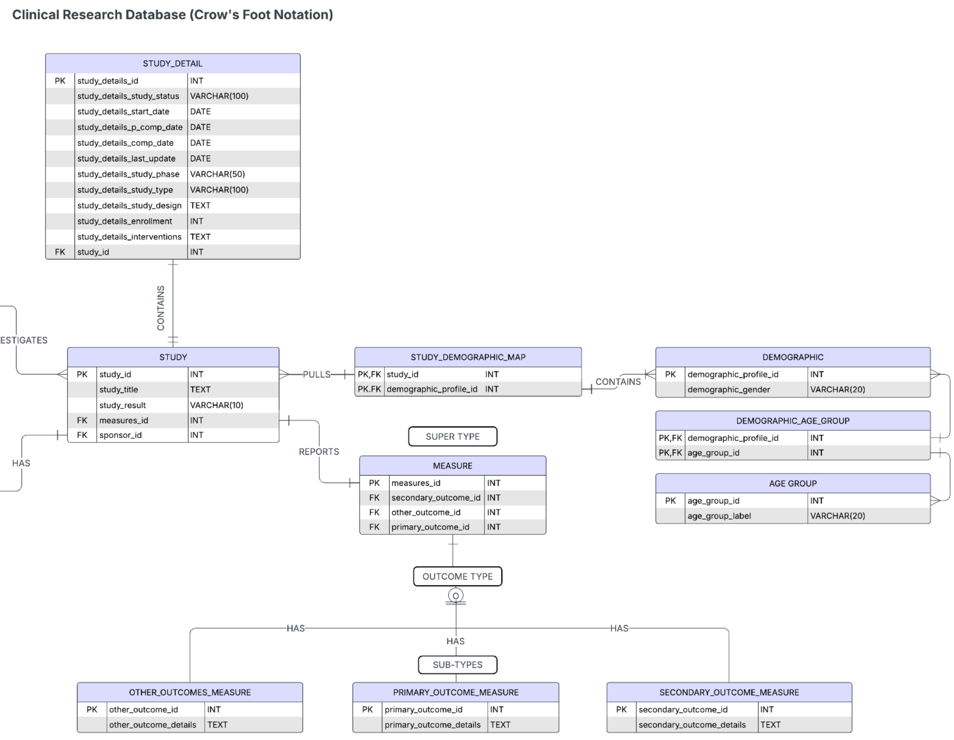
- This ensured participants can be easily queried by conditions
- CONDITION_ID values can be consistently reused across the STUDY

SPONSOR Table and COLLABORATOR Table (New) are linked together through a many-to-many relationship through the SPONSOR_COLLABORATOR (New) Junction Table.

SPONSOR_ID	COLLABORATOR_ID
201	301 (UCSF)
201	302 (Stanford)
202	302 (Stanford)

- This ensured participants can be easily queried by conditions
- COLLABORATOR_ID values can be consistently reused across SPONSOR.

Post-Normalized E-R Diagram


Clinical Research Database (Crow's Foot Notation)
STUDY_DETAIL
PK study_details_id INT
study_details_study_status VARCHAR(100)
study_details_start_date DATE
study_details_p_comp_date DATE
study_details_comp_date DATE
study_details_last_update DATE
study_details_study_phase VARCHAR(50)
study_details_study_type VARCHAR(100)
study_details_study_design TEXT
study_details_enrollment INT
study_details_interventions TEXT
FK study_id INT
CONTAINS
STUDY CONDITION
PK,FK condition_id INT
PK,FK study_id INT
CONDITION
PK condition_id INT
condition_name VARCHAR(255)
INVESTIGATES
STUDY
PK study_id INT
study_title TEXT
study_result VARCHAR(10)
FK measures_id INT
FK sponsor_id INT
PULLS
STUDY_DEMOGRAPHIC_MAP
PK,FK study_id INT
PK.FK demographic_profile_id INT
CONTAINS
DEMOGRAPHIC
PK demographic_profile_id INT
demographic_gender VARCHAR(20)
DEMOGRAPHIC_AGE_GROUP
PK,FK demographic_profile_id INT
PK,FK age_group_id INT
AGE GROUP
PK age_group_id INT
age_group_label VARCHAR(20)
HAS
REPORTS
SUPER TYPE
MEASURE
PK measures_id INT
FK secondary_outcome_id INT
FK other_outcome_id INT
FK primary_outcome_id INT
SPONSOR
PK sponser_id INT
sponsor_name VARCHAR(255)
sponsor_funder_type VARCHAR(100)
SPONSOR COLLABORATOR
PK,FK sponsor_id INT
PK,FK collabator_id VARCHAR(255)
COLLABORATOR
PK collaborator_id INT
collaborator_name TEXT
OUTCOME TYPE
HAS
HAS
HAS
SUB-TYPES
OTHER_OUTCOMES_MEASURE
PK other_outcome_id INT
other_outcome_details TEXT
PRIMARY_OUTCOME_MEASURE
PK primary_outcome_id INT
primary_outcome_details TEXT
SECONDARY_OUTCOME_MEASURE
PK secondary_outcome_id INT
secondary_outcome_details TEXT