**SYLLABUS**

**Course Name:** CLRE 255: Data Management & Informatics

**Course Instructor:** Raphael E. Cuomo, PhD ([racuomo@ucsd.edu](mailto:racuomo@ucsd.edu)). Office hours by appointment.

**Objectives:** By the end of the course, students will be able to:

* Understand and apply concepts of data management in a clinical research setting.
* Utilize SQL for data extraction, cleaning, optimization, and interpretation.

**Instructional Design:** The course will involve lectures, exercises, remote classes, and interactive discussions. Practical exercises will be given at the end of each week to provide hands-on experience. A typical class schedule will involve roughly one hour of lecture followed by one hour preparing, doing, and reviewing a SQL exercise.

**Textbook:** None

**Grading:** Grades will be assigned based on 100 points:

A = 94-100+ A- = 90-93 B+ = 87-89 B = 84-86 B- = 80-83 C = 70-79 D = 60-69 F < 59

**Evaluation:**

* **Weekly Exercises (20%):** There will be 8 exercises, each contributing 2.5% towards the final grade. These exercises are designed to provide hands-on experience with the concepts discussed in the weekly lectures. They will be graded based on completion, not on accuracy or correctness.
* **REDCap Lab (10%):** You willaccess REDCap and learn how to record clinical research data with this widely-used electronic data capture system.
* **Data Management Plan (35%):** You will need to develop a comprehensive data management plan for a hypothetical clinical study. The plan should include information about data collection, data cleaning, data storage, and data security, and it will be in the format of documents required for NIH grant submissions.
* **SQL Project (35%):** For this project, you will use the UCSD Health Research Database and conduct a study. You will need to use SQL to extract, manipulate, and analyze the data to answer these questions. This project will test your understanding of SQL as well as your ability to apply it in a practical context.

**REDCap Lab:** See handout.

**Data Management Plan:** Your data management plan should serve as a comprehensive outline for managing data throughout a hypothetical clinical study. It will be in the format of an NIH grant submission—specifically, the Data Management and Sharing Plan (DMSP), the DMSP Cost Justification, and the Data and Safety Monitoring Plan. You should include a description of the aims of the study and the types of data collected. The data collection section should elaborate on the tools and methodologies used to gather data effectively and accurately. In addition, include information about the data cleaning process, detailing both automated and manual procedures to ensure data reliability and accuracy. The plan should also provide detailed insights into the storage system and formats used to house the data securely and efficiently. It should detail data security measures, including technological strategies such as encryption and procedural methods such as access controls, that safeguard sensitive data against breaches and unauthorized access.

**SQL Project:** The task is to analyze the data and extract actionable insights to inform future research and clinical decisions. To initiate, you will formulate a research question that requires querying multiple tables in the database to resolve. For example, this could involve identifying the most commonly prescribed drug for a specific condition, determining the average patient age for a particular treatment, or finding correlations between certain conditions and drug exposures. You will then design and implement a comprehensive SQL querying strategy to extract the necessary data from the database. This process will involve data extraction, cleaning, transformation, and analysis using various SQL techniques taught throughout the course. If you have familiarity with R, it is preferred for you to include statistical testing code in R (though this is not required if you have not taken the program’s biostatistics courses). Please document each step in the process and provide the SQL queries (and R code, if applicable) used along with a brief explanation of each query's purpose. Once you've executed your queries and collected your data, you'll analyze the results and draw conclusions that answer your research question. This should not just be a simple reporting of the data but should involve analysis, trending, and correlation identification as appropriate.

**Academic Conduct:** All students are expected to abide by the university's policy on academic honesty and integrity. Plagiarism will not be tolerated.

**Attendance:** CREST/MAS program policy permits three absences without penalty. A fourth absence will result in a 15% reduction in your final grade. Be sure not to exceed 4 absences as this will result in automatic failure. **Redacted for FA23**.

**SCHEDULE**

**Week 1: Data in Clinical Research (Request Database Access & Set Up HS Duo by End of Day)**

* Lecture Topics: Review of syllabus, schedule, and canvas site. Introduction to SQL. Data types. Understanding the 3Vs: volume, velocity, and variety.
* *Technique for Exercise: SELECT, WHERE, COUNT, DISTINCT*

**Week 2: Introduction to Databases in Clinical Research**

* Lecture Topics: Design and structure of databases relevant to clinical research. Coded data and OMOP CDM.
* *Technique for Exercise: JOIN, Search strings with ILIKE, ORDER BY, GROUP BY*

**Week 3: Database Security and Privacy, and Data Governance (Remote Class with Mike Hogarth)**

* Lecture Topics: Techniques for data anonymization: masking, pseudonymization, and generalization. Symmetric and asymmetric encryption. Key principles of data governance: integrity, transparency, auditability, accountability, and stewardship. Steps for developing a data governance plan: defining objectives, scope, and responsibilities.
* Exhibit: Dr. Cuomo will showcase SAS code to transform a raw adverse event dataset to an ADaM compliant dataset for analysis

**Week 4: Database Maintenance and Optimization (Remote Class with Mike Hogarth)**

* Lecture Topics: Secure storage practices: access controls, backups, and physical security. Data security in the cloud. Encrypting data, managing access to data, optimization techniques.
* Exhibit: Dr. Cuomo will showcase a study that involves the use of both SQL and R in Databricks to obtain data and conduct analysis on the UC Health Data Warehouse (UCHDW)

**Week 5: Data Cleaning and Quality Assurance (DMP Due)**

* Lecture Topics: Importance of data quality: accuracy, completeness, consistency, timeliness, and validity. Structured and unstructured data in clinical research. Data cleaning techniques in clinical research, including the role of EDC systems and SQL commands.
* *Technique for Exercise: INNER JOIN, LEFT JOIN, OUTER JOIN*

**Week 6: Advanced Databases: Data Warehousing**

* Lecture Topics: Understanding the concept and function of data warehousing, including the use of SQL in data warehousing. Parallel processing.
* *Technique for Exercise: CTEs/Subqueries*

**Week 7: Data Integration and ETL Processes in Clinical Research (REDCap Lab Due)**

* Lecture Topics: Understanding the concept of data integration: homogenization of different data sources and types for unified viewing and analysis. The ETL process (Extract, Transform, Load): its role in clinical research data management, strategies, and common tools used.
* *Technique for Exercise: Window Functions*

**Week 8: Data Standards in Clinical Research and Normalization in Databases**

* Lecture Topics: Data standardization in clinical research, including the Clinical Data Interchange Standards Consortium (CDISC). Normalization, using standardized schemas.
* *Technique for Exercise: EXTRACT, HAVING, GROUP BY*

**Week 9: Compliance in Clinical Research Data Management**

* Lecture Topics: The role of a data steward in ensuring compliance (including GDPR data regulations), as well as upholding JEDI principles. Enforcing compliance through constraints and audits using SQL.
* *Technique for Exercise: UNION, CASE*

**Week 10: Data Interpretation (SQL Project Due)**

* Lecture Topics: Principles of data interpretation: contextualization, triangulation, and validation. SQL for contextualization, triangulation, and validation.
* *Technique for Exercise: COALESCE, WHERE EXISTS, SELF JOIN*