Applying Artificial Intelligence to Healthcare Research Design

Researchers familiar with the PICO framework (Population, Intervention, Comparison, and Outcome) can now leverage AI tools to streamline each step, from cohort selection in electronic health records (EHRs) to generating patient handouts, analyzing outcomes, and writing reports. This lab is designed as a hands-on workshop to demonstrate how AI can assist in transforming a traditional research workflow into a more efficient and reproducible process, using a simulated study of atorvastatin and exercise in patients with high LDL-cholesterol. Through practical exercises, participants will experience how AI can accelerate data preparation, support clinical trial planning, and simplify statistical analysis while ensuring research remains aligned with the principles of evidence-based medicine.

Objectives

By the end of this lab, participants will be able to:

- Understand how AI can be integrated into each step of the PICO framework.
- Use AI tools to clean and filter synthetic patient datasets according to inclusion/exclusion criteria.
- Apply AI for randomization, intervention assignment, and creation of patient-facing materials.
- Perform AI-assisted statistical comparison and generate visualization.
- Summarize outcomes and draft clinical interpretations with AI support.

Prerequisites

Participants are expected to have:

- Basic understanding of the PICO framework in clinical research.
- Familiarity with clinical concepts such as LDL-C, ALT, CK, and statin therapy.
- Introductory knowledge of research methodology and evidence-based practice.
- **Technical readiness:**
    - A computer (laptop or desktop) with internet access.
    - An active account with an **LLM/AI platform** (e.g., ChatGPT, Claude, Gemini, or equivalent).
    - Ability to open and work with simple **data files (CSV/Excel)**.
- (Optional) Basic experience with data files (CSV/Excel) and willingness to experiment with AI prompts or Python code generated by AI.

Evaluating LDL-C Reduction and Safety of Local vs Original Atorvastatin, With Exercise as an Adjunct, in Adults ≥35 Years: A Synthetic Cohort Study

Disclaimer

This lab uses synthetic data only. The dataset, results, and interpretations are created for educational and demonstration purposes. They do not represent real patients, real-world clinical outcomes, or validated medical evidence. The information in this workshop must not be used as a medical reference, clinical guideline, or basis for patient care.

Problem Statement

Cardiovascular disease remains a leading cause of morbidity and mortality, with elevated LDL-cholesterol (LDL-C) as a key modifiable risk factor. Although atorvastatin is widely used, many health systems consider switching from original (brand) products to locally manufactured generics to reduce cost. However, concise evidence on the **real-world lipid-lowering effectiveness and short-term safety** of local-made atorvastatin **at the common starting dose (10 mg)**—and how **exercise advice** may augment or modify its effect—is limited. Among adults **≥35 years** with **LDL-C at least 10% above guideline targets**, we lack a pragmatic comparison of: (1) LDL-C reduction and target attainment over **3 months** between **original vs local-made atorvastatin**, (2) the added benefit of **structured exercise advice vs no advice**, and (3) any **interaction** between drug type and exercise on efficacy and safety outcomes. This evidence gap constrains cost-effective treatment decisions and patient counseling in routine care.

Research Question

In adults ≥35 with high LDL-C (≥10% above target), does local-made atorvastatin 10 mg achieve non-inferior LDL-C reduction and similar short-term safety compared with original atorvastatin 10 mg, and is there additional benefit from providing exercise advice over 3 months?

Research Methodology

Study Design

A randomized, factorial, open-label, controlled trial will be conducted to compare the efficacy and safety of original vs locally manufactured atorvastatin (10 mg daily), with or without exercise advice, in adults with elevated LDL-C.

Population (P)

Inclusion criteria:

- Adults ≥35 years (male or female).
- LDL-C at least 10% above guideline target at baseline.

Exclusion criteria:

- Known contraindication to statins.
- History of liver disease, severe myopathy, or renal failure.
- Pregnancy or lactation.

Interventions (I)

Pharmacological intervention:

- Original atorvastatin 10 mg once daily.
- Local-made atorvastatin 10 mg once daily.

Lifestyle intervention:

- Structured exercise advice (moderate-intensity activity ≥150 min/week).
- No structured exercise advice (standard care).

Patients will be randomized into one of four groups (2 × 2 factorial design):

1. Original drug + No exercise advice
2. Original drug + Exercise advice
3. Local-made drug + No exercise advice
4. Local-made drug + Exercise advice

Comparison (C)

- Between original vs local-made atorvastatin groups.
- Between exercise vs no exercise advice groups.
- Interaction effects between drug type × exercise will also be assessed.

Outcomes (O)

Primary outcome:

- Mean change in LDL-C (mg/dL) from baseline to 3 months.

Secondary outcomes:

- Proportion of patients reaching LDL-C targets (<100 mg/dL for moderate risk; <70 mg/dL for high risk).
- Mean change in HDL-C and triglycerides.
- **Safety outcomes:** incidence of elevated liver enzymes (ALT), creatine kinase (CK), and muscle-related symptoms.

Sample Size & Randomization

- A total of N participants will be equally randomized into 4 groups using a computer-generated block randomization method.
- Allocation ratio: 1:1:1:1.
- Sample size is designed to detect non-inferiority of local-made atorvastatin within a pre-specified margin (e.g., 5 mg/dL difference in LDL-C reduction).

Data Collection

- Baseline assessment: demographics, risk factors, laboratory values (LDL-C, HDL-C, triglycerides, ALT, CK).
- Follow-up at 3 months: repeat laboratory measures, adherence check, adverse event reporting.
- Exercise adherence will be assessed via self-report.

Data Analysis

- Descriptive statistics for baseline characteristics.
- Primary analysis: ANCOVA model with follow-up LDL-C as dependent variable, baseline LDL-C as covariate, and drug type, exercise advice, and interaction as independent factors.

- Secondary analysis: Logistic regression for LDL-C target attainment; chi-square/Fisher's exact test for safety outcomes.
- Non-inferiority margin: 5 mg/dL LDL-C reduction difference between local and original atorvastatin.

Ethical Considerations

- The study will be conducted according to the Declaration of Helsinki and Good Clinical Practice (GCP).
- Informed consent will be obtained from all participants.
- Data confidentiality will be maintained throughout.

PICO Step	How AI Helps	Example in The Demo
P – Population	AI filters cohorts from raw HIS/EHR data; detects missing/invalid entries; enforces inclusion/exclusion rules.	From 1000 HIS mixed records → AI selects 684 eligible patients (≥35 yrs, LDL ≥10% above target, consent signed, no pregnancy, no statin allergy).
I – Intervention	AI randomizes patients, balances groups, generates patient instructions & survey forms automatical.	2×2 factorial randomization → 4 arms (Original vs Local drug × Exercise vs No Exercise). AI generated patient handouts + survey templates.
C – Comparison	AI merges follow-up data with baseline, handles missing visits, and runs advanced models (ANCOVA, regression).	Merged visit 12-week results → AI showed exercise lowers LDL ~10 mg/dL more; drug type had no significant difference.
O – Outcome	AI produces summary tables, adjusted mean plots, and plain-language interpretations for researchers/clinicians.	Outcome summary table + visualization; AI explains: "Exercise is dominant factor, drug type effect negligible."

Lab 1 - Population

AI filters cohorts from raw HIS/EHR data; detects missing/invalid entries; enforces inclusion/exclusion rules. From 1000 HIS mixed records → AI selects 684 eligible patients (≥35 yrs, LDL ≥10% above target, consent signed, no pregnancy, no statin allergy).

File: Download https://bhattaraprot.com/rawfiles.zip
Input File : synthetic_HIS_1000_mixed.csv

Prompt Example 1 – Natural Language

You are a clinical data assistant.
From this CSV file of hospital patients, select those eligible for our study.
Inclusion criteria:
- Age ≥ 35 years
- LDL cholesterol ≥ 10% above guideline threshold
- Consent_sign = Y

Exclusion criteria:
- Pregnancy_flag = 1
- Statin_allergy_flag = 1
- Diagnosis includes liver_disease, myopathy, or renal_failure

Return a clean table with only eligible patients and keep only these columns:
patient_id, age, sex, ldl, hdl, alt, ck, diagnosis_codes, consent_sign

Prompt Example 2 – Python Script

Write a Python script in pandas to filter a dataset named synthetic_HIS_1000_mixed.csv.
Apply these rules:
- Include patients age ≥35 and LDL ≥ threshold (10% above guideline).
- Exclude pregnancy_flag=1, statin_allergy_flag=1, or diagnosis_codes containing ["liver_disease","myopathy","renal_failure"].
- Keep only these columns: patient_id, age, sex, ldl, hdl, alt, ck, diagnosis_codes, consent_sign.
Save result to study_cohort_clean.csv.

Lab 2: Intervention

Assign patients to study arms, generate study materials (handouts, surveys), and plan follow-ups.

Prompt Example 1 – Natural Language (Randomization)

You are a clinical trial coordinator.
We have a dataset of eligible patients in study_cohort_clean.csv.
Randomize patients into 4 arms equally (2×2 factorial):

1. Original Drug + No Exercise
2. Original Drug + Exercise
3. Local Drug + No Exercise
4. Local Drug + Exercise

Add new columns: assigned_drug, exercise_advice, random_group.
Export result as study_cohort_randomized.csv

Prompt Example 2 – Natural Language (Patient Handouts)

You are a health educator.
Write a simple patient handout (1 page, plain Thai language) for each intervention group:

1. Original Drug + No Exercise
2. Original Drug + Exercise
3. Local Drug + No Exercise
4. Local Drug + Exercise

Include:
- What medicine they receive
- If they should exercise or not (and what kind of exercise)
- Reminder to return for blood tests at week 6 and week 12

Prompt Example 3 – Micro-Survey for Patients

Generate a 4-question survey (Google Form format) for patients after 12 weeks:
- Did you take your medicine regularly? (Yes/No)
- How many days per week did you exercise? (0–7)
- Did you experience any muscle pain? (None/Mild/Severe)
- Did you experience nausea or liver symptoms? (Yes/No)

Prompt Example 4 – Python Script (Randomization & Appointments)

Write a Python script to:
1. Load study_cohort_clean.csv
2. Randomly assign each patient to one of 4 groups (Original vs Local × Exercise vs No).
3. Generate a follow-up schedule: baseline date = today, plus week 6 and week 12 visits.
4. Save to study_cohort_randomized_with_dates.csv

Lab 3: Comparison

Merge baseline + follow-up, analyze treatment effects, and visualize results.

Files: Download https://bhattaraprot.com/package.zip , https://bhattaraprot.com/rawfiles.zip
Input: visit12_response_filled.csv (demo) or visit12_respons.csv (your data)

Prompt Example 1 – Natural Language (Data Merge)

You are a data analyst.
We have 2 CSVs:
- study_cohort_randomized_with_dates.csv (baseline + allocation)
- visit12_response.csv (12-week outcomes)

Merge them by patient_id.
Keep baseline LDL, follow-up LDL, drug type, and exercise assignment.
Prepare an analysis-ready table called analysis_ready.csv

Prompt Example 2 – Natural Language (Statistical Model)

Perform an ANCOVA analysis:
Outcome = follow-up LDL
Covariate = baseline LDL
Factors = assigned drug (original vs local), exercise (yes vs no), and their interaction.

Report adjusted means and p-values for each group.
Interpret results in plain language for clinicians.

Prompt Example 3 – Natural Language (Visualization)

Create a boxplot or adjusted means chart showing follow-up LDL across the 4 study groups:
- Original + No Exercise
- Original + Exercise
- Local + No Exercise
- Local + Exercise

Highlight whether exercise or drug type has the stronger effect.

Prompt Example 4 – Python Script (Full Analysis)

Write a Python script using pandas and statsmodels to:
1. Merge baseline cohort with 12-week outcomes.
2. Fit an ANCOVA model: followup_ldl ~ baseline_ldl + drug + exercise + drug*exercise.
3. Output adjusted means for each group.
4. Save analysis results to analysis_results.csv.
5. Create a visualization (boxplot) of LDL by group.

Lab 4: Outcome

Summarize results, interpret them for clinicians, and prepare outputs for reporting.

Prompt Example 1 – Natural Language (Summary Table)

You are a medical research assistant.
From analysis_ready.csv, create an outcome summary table with:
- Group (drug × exercise)
- N patients
- Mean baseline LDL
- Mean follow-up LDL
- Mean LDL change
- Mean ALT and CK
- Adherence rate
- Adverse event rate

Format the table for publication (rounded, clear labels).

Prompt Example 2 – Natural Language (Interpretation)

Interpret the results of our study in plain clinical language:

- Which factor had the strongest effect on LDL reduction?
- Did original vs local drug show meaningful differences?
- Were safety markers (ALT, CK) acceptable?
- Summarize in 3–4 sentences for physicians.

Prompt Example 3 – Natural Language (Report Generation)

Draft the Results section of a research report using the analysis results.
Include:
- Main findings on LDL changes
- Safety observations
- Adherence/AEs
- Clinical implication (exercise importance, drug equivalence)

Make it suitable for a journal submission.

Prompt Example 4 – Python Script (Visualization & Export)

Write a Python script to:
1. Read analysis_results.csv
2. Generate a bar chart of mean LDL reduction by group
3. Generate a line chart of LDL change from baseline to follow-up for each group
4. Save all plots as PNG files
5. Export summary stats into outcome_summary.csv