MSc Health Data Science

LSHTM_2487_2022 – Health Data Management

Assignment 2022-23

Aim

This assignment assesses your ability to understand an operational question, extract relevant data, perform basic transformations on that data, summarize the results to answer the question in text, tables, and visualisations, and provide auditable documentation of your work. You will use the MIMIC III database as the data source, and any analyses and visualisations should be performed in a combination of SQL (in the PostgreSQL syntax) and R or Python or Excel.

Assessment criteria

You will be assessed on 3 products: an executive report, SQL query file(s) to extract data from MIMIC III covering the necessary information for the report, and code file(s) to transform the extracted data. The report will be evaluated for i) responsiveness to the requests, ii) accuracy, iii) clarity and organization. The SQL will be evaluated for i) correctness, ii) auditability, iii) performance. The analysis code will be evaluated for i) correctness, ii) auditability, iii) reproducibility. Since the SQL and analysis code work in tandem, these will be evaluated as complementary parts.

Task instructions

You are a data analyst for Beth Israel Deaconess Medical Center, where patient data is captured in the MIMIC III database. Hospital management identified worrying signs associated with a patient, and wants to review trends in patient data. You are tasked to produce an executive summary report; per normal audit policy, you need to record and document your data extraction and analysis.

Hospital policy is to minimise any data extracted from the database for privacy reasons, even though that data is deidentified, so ensure queries are appropriately filtered and select the minimum columns to respond to the report needs.

Your report should:

i) Provide summary data for patient 42130 (demographic data, total stay time, total time in the ICU if any, care units and wards, all diagnoses with ICD9 codes, and any prescriptions) for the most recent visit where Simvastatin was prescribed. This data should be captured in table, with, if necessary, limited text explanation.

ii) Provide a time series visualisation of routine quantitative vital signs for patient 42130 for that stay. If necessary for interpretation, provide a caption.

iii) Provide a summary table for admissions of similarly aged patients (60-65 years old), also with a cardiac device (ICD9 codes starting with V450) and stay time in the ICU, capturing total stay time in terms of median, 25/75% quantiles, and min-max, and total ICU time per admission (same ranges, central indicator). Stratify the summaries by gender and whether or not the patients died in ICU (defined as time of death within a 6 hour window before and after any ICU stay).

iv) For those patients that died, provide a table summarising their stay that corresponded with death. By patient id, list first care unit, gender, age at admission, total days in ICU, and all ICD9 diagnoses and codes associated with ICU periods.

v) Provide a visualisation comparing for each ICU: average total ICU stay time per admission for all patients, and then same average for patient subsets corresponding to a) all patients aged 60-65 with a cardiac device, and b) with and c) without simvastatin prescribed. The visualisation should also show these values aggregated across all units to aid in comparison.

vi) If necessary, summarise in a notes section the decisions you made as an analyst to resolve any ambiguities in the above guidance. Ensure all elements of the report indicate what they are, for context and posterity.

The report should be submitted as pdf, maximum two pages. The audit elements should provided as plain text file(s) for all SQL queries and plain text files for any analysis code (R, Python, or mixed code / report generation formats like Rmarkdown) or an xlsx file if that was used for analysis and figure generation. While there are no specific limits for size of the SQL and analysis code files, keep in mind the requirement for reasonable audit effort and strive to balance clear and concise in your code.

Notes

1) In general, “auditability” means that a technically adept reviewer can confirm the intent and correctness of your code with reasonable effort, and reproduce your analysis precisely if necessary with minimal effort. As such, code should be documented for clarity as appropriate, and recorded in way that makes its use easy.
2) No clinical or statistical analysis beyond directed calculations of any data is required.
3) You may need to write multiple queries to answer a particular question. In some cases, obtaining the relevant data could be accomplished with a single query with complex joins or by a series of queries where you manually substitute results from one query into a filtering clause for subsequent one. You should balance auditability of the SQL with desire to minimize patient data that resides outside the data base.

Submission Deadline

Assignments must be uploaded to the Assignment submission point on the HDM Moodle pages no later than **17:00 (UK time) on Mon 14th Nov 2022.**

Assessment criteria

- Responsiveness, correctness, clarity, and directness of your report (1/3 of this assignment, or 25% of overall final assessment including MCQ)
- Correctness, auditability (e.g., layout, comments, approach to queries), and performance (e.g. extraction of minimal data necessary, query design) in your SQL (1/3 of this assignment)
- Correctness, auditability, and performance of the analysis code (1/3 of this assignment)