

Course/Data: bộ dữ liệu khóa học FLUID-ICU

Cấp độ 1 dùng bộ dữ liệu này để phân tích và viết bản thảo trong Tuần 5-8. Cấp độ 2 dùng lại bộ dữ liệu này để thực hành.

> **DỮ LIỆU MÔ PHỎNG: KHÔNG PHẢI NGƯỜI BỆNH THẬT.** FLUID-ICU là một nghiên cứu hư cấu. Mọi giá trị đều do một chương trình máy tính tạo ra với hạt giống ngẫu nhiên (random seed) cố định. Dữ liệu không mô tả bất kỳ người bệnh, bệnh viện, ICU hay quốc gia thật nào. Bản thảo nào viết từ bộ dữ liệu này cũng phải ghi rõ dữ liệu là mô phỏng và không được nộp như một nghiên cứu thật.

FLUID-ICU là một nghiên cứu thuần tập tiến cứu đa trung tâm mô phỏng, gồm **620 người lớn bị nhiễm khuẩn huyết** nhập 4 ICU ("ICU A" đến "ICU D") từ tháng 1/2023 đến tháng 12/2024. Bộ dữ liệu đưa ví dụ xuyên suốt của khóa học vào Buổi 5-8: *cân bằng dịch tích lũy dương vào ngày 3 có liên quan với tử vong 28 ngày cao hơn không?* Dữ liệu được xây dựng để các bài học kinh điển hiện ra trong kết quả: nhiễu do chỉ định, dữ liệu thiếu, một kết cục phân phối lệch, và một tương tác đáng bàn luận. Các khoảng giá trị lâm sàng dựa theo hai bài báo của khóa học (Boulet 2024; Hyun 2023).

Thư mục này có gì

Tệp / thư mục	Nội dung	Dùng ở
fluid_icu_raw.csv, fluid_icu_raw.xlsx	Dữ liệu đúng như 4 ICU "gửi về": 623 dòng với các lỗi sát thực tế	Trình diễn làm sạch ở B5
fluid_icu_clean.csv, fluid_icu_clean.rds	Dữ liệu sẵn sàng để phân tích: 620 người bệnh, 32 biến (tệp .rds giữ các mức factor của R)	Mọi phân tích
data_dictionary.md, data_dictionary.csv	Mọi biến số: nhãn, loại, đơn vị/mã, khoảng cho phép, mã giá trị thiếu, ghi chú	B5
cleaning_log.md	Mọi vấn đề tìm thấy trong tệp thô và cách đã xử lý, viết bằng lời đơn giản	B5
SAP_FLUID-ICU.md	Kế hoạch phân tích thống kê, viết "trước khi phân tích" (14 mục theo mẫu SAP của khóa học)	B6, B7
key_results.md	Mọi con số được trích trong slide và tài liệu phát tay, kèm script nguồn	Giảng viên
results_packs/Q1 ... Q5	Mỗi câu hỏi một gói (Q1 đến Q5): results_pack_QN.md, tables/ (CSV + Word) và figures/ (PNG)	B5-B8: cả lớp cùng đọc Q1; mỗi học viên viết một bản thảo về một câu hỏi tự chọn
results_packs/Results_Pack_QN.docx	Cùng gói kết quả đó dưới dạng một tài liệu Word (chia sẻ dạng tệp hoặc PDF)	Học viên B5-B8
R/	Các script (xem bên dưới)	Trình diễn của giảng viên
python/make_slide_figures.py	Vẽ các hình cho slide bằng tiếng Anh và tiếng Việt vào Course/Resources/figures/{en,vi}/data/	Dựng slide

Tệp / thư mục	Nội dung	Dùng ở
output/	Những gì các script trình diễn ghi ra: tables/, figures/, slide_data/ và fluid_icu_clean_from_raw.csv	Trình diễn
FLUID-ICU.Rproj	Mở tệp này trong RStudio: thư mục làm việc được đặt là Course/Data	Trình diễn

(B5 = Buổi 5, tương tự cho các buổi khác.)

Năm câu hỏi Q1 đến Q5 (mục 1 của bản thiết kế khóa học)

Mỗi học viên chọn một câu hỏi cho thảo luận của mình. Trên lớp, giảng viên cùng cả lớp làm qua các câu hỏi (Q1 đầy đủ, Q2 đến Q5 dưới dạng ví dụ mẫu ngắn hơn).

Q	Câu hỏi	Phương pháp chính	Gói
Q1	Cân bằng dịch dương vào ngày 3 và tử vong 28 ngày (câu hỏi chính)	Hồi quy logistic (thô và hiệu chỉnh), Kaplan-Meier, Cox, tương tác với suy tim	Q1
Q2	Cân bằng dịch tích lũy và tổn thương thận cấp	Hồi quy logistic, cân bằng dịch tính theo mỗi 10 ml/kg	Q2
Q3	Cân bằng dịch tích lũy và số ngày không thở máy	Mann-Whitney; hồi quy tuyến tính	Q3
Q4	Lactat lúc nhập viện và tử vong 28 ngày; ảnh hưởng của lactat bị thiếu	Phân tích trường hợp đầy đủ so với quy gán đa bội (mice, m = 20)	Q4
Q5	Cân bằng dịch và thời gian nằm ICU ở người sống	Trung vị (IQR); Mann-Whitney; hồi quy tuyến tính trên thang log	Q5

Cách chạy các script

Mở FLUID-ICU.Rproj trong RStudio (hoặc đặt thư mục làm việc là Course/Data) và chạy các script theo thứ tự. Từ cửa sổ dòng lệnh:

```
cd Course/Data
Rscript R/00_generate_fluid_icu.R # (course team only) simulate; writes raw + clean files
Rscript R/01_clean.R # S5 demo: raw -> clean; confirms identical to the clean file
Rscript R/02_describe.R # S5 demo: Table 1 + histogram, boxplot, bar chart
Rscript R/03_compare.R # S6 demo: t-test, Mann-Whitney, chi-square
Rscript R/04_models.R # S6 demo: crude vs adjusted OR, forest plot, KM, Cox, interaction
Rscript R/05_missing.R # S6 demo: missing lactate, complete case vs multiple imputation
Rscript R/06_results_packs.R # (course team) results packs, key_results.md, slide data (~1 min)
python python/make_slide_figures.py # (course team) slide figures EN + VN (needs Deck_and_Build/figures)
```

Ý nghĩa các dòng trên:

- 00: mô phỏng dữ liệu và ghi tệp thô và tệp sạch (chỉ nhóm biên soạn chạy).
- 01: làm sạch từ thô sang sạch và xác nhận kết quả giống hệt tệp sạch (trình diễn B5).
- 02: Bảng 1, biểu đồ tần suất, biểu đồ hộp, biểu đồ cột (trình diễn B5).
- 03: kiểm định t, Mann-Whitney, khi bình phương (trình diễn B6).

- 04: OR thô và hiệu chỉnh, biểu đồ rừng, Kaplan-Meier, Cox, tương tác (trình diễn B6).
- 05: lactat bị thiếu, phân tích trường hợp đầy đủ so với quy gán đa bội (trình diễn B6).
- 06: gói kết quả, `key_results.md` và dữ liệu cho slide, khoảng 1 phút (nhóm biên soạn).
- Script Python: hình cho slide bằng tiếng Anh và tiếng Việt (nhóm biên soạn).

Học viên không bao giờ cần chạy gì: học viên đọc các kết quả được trình chiếu trong phần trình diễn trực tiếp (chia sẻ dưới dạng PDF) và gói kết quả cho câu hỏi của mình. Các script trình diễn được viết để chạy **từng khối một** trong RStudio (Ctrl+Enter), mỗi khối có một dòng chú thích phía trước nói cần chỉ ra điều gì. Các script 01-05 chỉ đọc tệp dữ liệu và ghi vào `output/`. Script 00 ghi đề các tệp dữ liệu (với kết quả giống nhau mỗi lần chạy, vì hạt giống ngẫu nhiên cố định).

Các gói R cần có

R 4.3 trở lên (đã dựng và kiểm tra với R 4.6.1).

```
install.packages(c("tidyverse", "janitor", "writexl", "readxl", "gtsummary", "gt",
                  "flextable", "officer", "broom", "survival", "survminer", "mice"))
```

Python (chỉ để vẽ hình cho slide): Python 3.10 trở lên với `matplotlib`, `numpy`, `pandas`; phông chữ Century Gothic (tiếng Anh) và Segoe UI (tiếng Việt).

Những điều cần biết trước khi giảng

- **Tái lập chính xác.** `01_clean.R` kết thúc bằng bước kiểm tra kết quả của nó trùng khớp với `fluid_icu_clean.csv` từng giá trị và từng byte. Nếu script in ra `FALSE`, tệp thô hoặc tệp sạch đã bị sửa bằng tay.
- **Các con số.** `key_results.md` liệt kê mọi con số được trích dẫn. Nếu chạy lại `00_generate_fluid_icu.R` trên một phiên bản R khác, về nguyên tắc các số ngẫu nhiên có thể khác. Khi đó hãy chạy lại 01-06 và script Python, rồi kiểm tra lại `key_results.md`.
- **Mô phỏng nghĩa là mô phỏng.** "Tác động" của dịch được cố ý đưa vào dữ liệu (chủ yếu ở người bệnh suy tim) cùng với nhiễu mạnh do mức độ nặng. Dữ liệu này không thể cho ai biết dịch có gây hại cho người bệnh thật hay không.