

Buổi 5 (Trình diễn trực tiếp): làm sạch FLUID-ICU và lập Bảng 1 bằng R

Nền tảng Nghiên cứu Lâm sàng · Tuần 5, Thứ Bảy 19:30 đến 22:30 (trực tuyến) · Neudata · kịch bản cho giảng viên

	Phần A: làm sạch	Phần B: Bảng 1 và biểu đồ
Thời điểm	5.2 Làm quen bộ dữ liệu khóa học (khoảng 20:25)	5.4 Mô tả bộ dữ liệu khóa học (khoảng 21:31)
Thời lượng	8 phút	6 phút
Script	Course/Data/R/01_clean.R	Course/Data/R/02_describe.R
Học viên	theo dõi màn hình chia sẻ của giảng viên và dự đoán trong khung chat (học viên không chạy R)	như trên

> **FLUID-ICU là dữ liệu MÔ PHỎNG**. Nhắc lại điều này mỗi lần mở dữ liệu: không có người bệnh, bệnh viện hay quốc gia thật nào. Mọi con số dưới đây lấy từ `Course/Data/key_results.md` và `Course/Data/cleaning_log.md`.

Trước buổi học (tối hôm trước và 10 phút trước giờ học)

- [] Đã cài R (4.x) và RStudio trên máy tính dùng để chia sẻ màn hình; các gói: tidyverse, janitor, gtsummary, flextable, survival, survminer, mice, broom.
- [] Mở `Course/Data/FLUID-ICU.Rproj` (đặt thư mục làm việc). Chạy trọn `01_clean.R` và `02_describe.R` một lần từ đầu đến cuối. Cả hai phải chạy xong không lỗi.
- [] Tăng cỡ chữ của cửa sổ soạn thảo và cửa sổ console (Tools, Global Options, Appearance: 18 đến 20 pt) để học viên đọc được trên màn hình nhỏ. Chỉ chia sẻ cửa sổ RStudio.
- [] Mở sẵn các hình dự phòng trong trình xem ảnh: `Course/Resources/figures/vi/data/(raw_vs_clean.png, table1_snapshot.png, fluid_balance_hist.png, sofa_by_group.png, los_skewed.png; bản tiếng Anh ở Course/Resources/figures/en/data/)` và `Course/Data/output/figures/`.
- [] Mở `Course/Data/cleaning_log.md` và `data_dictionary.md` trong một thẻ trình duyệt: giảng viên sẽ cho lớp xem hai tệp này.

Quy tắc khi trình diễn: chạy từng khối một (Ctrl+Enter). Trước mỗi khối, hỏi cả lớp *"anh chị dự đoán sẽ thấy gì?"* và đọc hai hoặc ba câu trả lời trong khung chat. Sau mỗi khối, nói *vấn đề nào vừa được sửa và việc sửa đó được ghi vào nhật ký*. Không bao giờ gõ code mới trực tiếp, vì các script đã hoàn chỉnh.

Phần A: Từ dữ liệu thô đến dữ liệu sạch (01_clean.R, 8 phút)

Bước (khối trong script)	Cho lớp xem	Kết quả dự kiến	Nói
0. Đọc dữ liệu thô dạng văn bản	<code>dim(raw)</code> , <code>glimpse(raw)</code>	623 dòng , 29 cột, tất cả là <code><chr></code>	"Nghiên cứu tuyển 620 người bệnh. Vì sao lại có 623 dòng?"

Bước (khối trong script)	Cho lớp xem	Kết quả dự kiến	Nói
1. Tên cột	<code>names (dat)</code>	tên ngắn, chữ thường (<code>map_mmhg</code> , <code>vent_free_days_28</code>)	"Tên không có dấu cách: giống nhau trong mọi script."
2. Bản ghi trùng	<code>get_dupes (study_id)</code>	3 mã nghiên cứu xuất hiện hai lần, các dòng giống hệt nhau; sau <code>distinct ()</code> còn 620	"Mỗi người bệnh chỉ được đếm một lần."
3. Đơn vị ICU	<code>count (site)</code> trước/sau	các biến thể như <code>icu a</code> , <code>ICU-A</code> : gộp thành 4 ICU	"Nếu không, R sẽ coi đó là nhiều ICU khác nhau."
4. Ngày tháng	ba định dạng	2023-03-14, 14/03/2023, 14-Mar-2023: thành một kiểu ngày; 0 lỗi chuyển đổi	"03/04 là tháng Ba hay tháng Tư? Không bao giờ đoán."
5-6. Giới tính và có/không	<code>tabyl ()</code> trước/sau	nhiều cách viết (M, m, male, F, f, "Female"; Y, 1, yes) thành Male/Female, Yes/No	"1 và Yes phải có cùng một nghĩa."
7-8. Số và mã	<code>summary ()</code> của lactat, MAP, APACHE II	giá trị lớn nhất 999 , nhỏ nhất -99 : đặt là NA	"Một giá trị lactat 999 sẽ làm hỏng số trung bình."
9. Đơn vị đo	trung vị lactat theo ICU	ICU C cao gấp khoảng 9 lần (mg/dL): chia cho 9.01	"Chỉ phát hiện được vì ta tóm tắt theo từng ICU. Hỏi lại phòng xét nghiệm của ICU đó."
10. Giá trị vô lý	kiểm tra khoảng giá trị	tuổi 210, cân nặng 7 kg, lượng dịch ra âm: đặt là NA và gửi câu hỏi xác minh; vẫn giữ người bệnh	"Ta không thể đoán là 21 hay 71. Không bao giờ xóa người bệnh."
12. Biến dẫn xuất	công thức cân bằng dịch	tính <code>cum_fluid_balan ce_d3_ml_kg,positi ve_balance_d3, aki_any</code>	"Giá trị dẫn xuất được tính ra, không bao giờ gõ tay."
13. Phần còn lại	<code>colSums (is.na (dat))</code>	lactat thiếu 93, rải rác vài giá trị thiếu ở biến khác	"Đây là những khoảng trống THẬT. Ta xử lý chúng ở Buổi 6."
14. Tính tái lập	lệnh <code>cat ()</code> cuối cùng	<code>Cleaned data identical to the reference clean file: TRUE/ byte-for-byte identical: TRUE</code>	"Ai cũng có thể chạy lại và được đúng dữ liệu này."

Nếu thiếu thời gian: chỉ chạy các Bước 0, 2, 9, 10 và 14 (4 phút) và cho xem nhật ký làm sạch cho phần còn lại.

Kết thúc Phần A bằng cách mở `cleaning_log.md`: *"Mỗi khối anh chị vừa thấy là một dòng ở đây: vấn đề, bao nhiêu trường hợp, ta đã làm gì, vì sao."* Sau đó chiếu slide Từ 623 dòng đến 527 trường hợp đầy đủ (620 người được tuyển, rồi 618 người có cân bằng dịch ngày 3, rồi 527 trường hợp đầy đủ).

Phần B: Bảng 1 và biểu đồ (02_describe.R, 6 phút)

Khối	Cho lớp xem	Kết quả dự kiến (key_results.md)	Nói
1. Nạp dữ liệu	<code>dim(icu), count(positive_balance_d3)</code>	620 người bệnh, cân bằng dịch dương 373/618 (60.4%) , 2 người không có giá trị cân bằng dịch	"618, không phải 620: hai người bệnh không tính được cân bằng dịch."
2. Xem hình dạng trước	trung bình/SD của tuổi, trung vị/IQR của lactat; biểu đồ tần suất nhanh	tuổi trung bình 62.6 (SD 15.2) , lactat lệch	"Đối xứng: trung bình (SD). Lệch: trung vị (IQR)."
3. Bảng 1	bảng gtsummary theo nhóm	nhóm âm n = 245, nhóm dương n = 373; SOFA 6 (4 đến 8) so với 8 (6 đến 10) ; thuốc vận mạch 24.9% so với 55.0% ; thở máy 39.6% so với 57.9% ; lactat 2.2 so với 2.9 mmol/L; suy tim 22.0% so với 11.5% ; có hiển thị số thiếu; không có giá trị p	"Lúc ban đầu các nhóm có giống nhau không? Không: nhóm dương nặng hơn."
4. Biểu đồ tần suất	cân bằng dịch theo nhóm	trung vị 8.8 ml/kg (IQR -10.3 đến 26.2)	"Phơi nhiễm liên tục, cắt tại 0."
5. Biểu đồ hộp	SOFA theo nhóm, có chấm điểm	trung vị 6 so với 8	"Cho thấy từng người bệnh."
6. Biểu đồ cột	nguồn nhiễm khuẩn, đã sắp xếp	hô hấp thường gặp nhất	"Biến phân loại: cột đã sắp xếp, không dùng biểu đồ tròn 3-D."
7. Kết cục	mô tả, chưa so sánh	tử vong 28 ngày 179/620 (28.9%) , thời gian nằm ICU trung bình 7.3 so với trung vị 6.0 ngày	"Hôm nay mô tả, thứ Bảy tới mới so sánh."

Kết thúc Phần B: "Bảng 1 này có trong mọi gói kết quả, từ Q1 đến Q5. Giờ chúng ta cùng đọc nó như một người phản biện." Sau đó chuyển sang phần cả lớp cùng đọc Bảng 1 (Bài tập 6).

Kế hoạch dự phòng

Sự cố	Cách xử lý
R hoặc một gói bị lỗi	Chiếu <code>raw_vs_clean.png</code> và giải thích các vấn đề (Phần A); chiếu <code>table1_snapshot.png</code> , <code>fluid_balance_hist.png</code> , <code>sofa_by_group.png</code> (Phần B). Các ý giảng dạy giữ nguyên.
Chia sẻ màn hình không hiển thị RStudio đủ rõ	Dùng các hình PNG và <code>cleaning_log.md</code> trong trình duyệt với mức thu phóng 150%.
Bị trễ giờ	Phần A: Bước 0, 2, 9, 10, 14. Phần B: chỉ khối 3 và 5.
Một học viên muốn tự chạy	Giới thiệu <code>Course/Data/README.md</code> (bản tiếng Việt: <code>Course/Data/README_VN.md</code>) và khóa học miễn phí <i>Clinical Data Analysis in R</i> , và đề nghị hỗ trợ sau buổi học.

Các câu hỏi có thể gặp

"Đây có phải dữ liệu thật không?" Không. Đây là dữ liệu mô phỏng, được thiết kế để giống dữ liệu ICU và để minh họa các vấn đề kinh điển. Bản thảo nào dùng dữ liệu này cũng phải nói rõ điều đó.

"Sao không sửa lỗi trong Excel?" Sửa tay không để lại dấu vết và không lặp lại được, và một lần dán nhầm thì không ai thấy. Script kèm nhật ký thì người phản biện kiểm tra được và ai cũng chạy lại được.

"Sao đặt tuổi 210 là thiếu thay vì đoán là 21?" Có thể là 21, 71 hay 10, và đoán là tự tạo ra dữ liệu. Đặt là thiếu, hỏi lại ICU đó, và chỉ sửa khi có tài liệu gốc.

"Sao Bảng 1 không có giá trị p?" Bảng 1 mô tả các nhóm. SAP ghi rõ điều này. Trong thử nghiệm ngẫu nhiên có đối chứng, mọi khác biệt đều do ngẫu nhiên. Trong nghiên cứu thuần tập, khác biệt là yếu tố gây nhiễu cần hiệu chỉnh, không phải để kiểm định.