

Nhật ký làm sạch dữ liệu FLUID-ICU

> DỮ LIỆU MÔ PHỎNG: KHÔNG PHẢI NGƯỜI BỆNH THẬT. Nhật ký này thuộc bộ dữ liệu giảng dạy FLUID-ICU.

Tệp thô	fluid_icu_raw.csv (nội dung giống hệt trong fluid_icu_raw.xlsx): 623 dòng, 29 cột, đúng như nhận từ ICU A-D
Tệp sạch	fluid_icu_clean.csv / .rds: 620 người bệnh, 32 biến
Script làm sạch	R/01_clean.R (mỗi bước dưới đây là một khối lệnh của script)
Kiểm tra	Script dựng lại tệp sạch từ tệp thô và xác nhận kết quả giống hệt, từng giá trị và từng byte
Quy tắc	Không bao giờ sửa tệp thô. Mọi thay đổi đều do script thực hiện và được ghi lại ở đây.

Các vấn đề phát hiện và cách đã xử lý

Bước	Vấn đề phát hiện	Số lượng	Cách xử lý	Lý do
1	Tên cột gõ theo kiểu bảng tính: có dấu cách, dấu ngoặc, chữ hoa (Weight (kg), MAP (mmHg), VFD 28)	29 cột	Đổi thành tên ngắn, chữ thường bằng clean_names(); đổi tên 4 cột (ví dụ map_mm_hg thành map_mmhg, vfd_28 thành vent_free_days_28)	Tên không có dấu cách thì dễ gõ và giống nhau trong mọi script
2	Bản ghi trùng: cùng một người bệnh được nhập hai lần	3 mã (FLU-139, FLU-274, FLU-574): 623 dòng cho 620 người bệnh	Kiểm tra hai bản giống hệt nhau ở mọi cột, rồi giữ lại một bản	Mỗi người bệnh chỉ được đếm một lần
3	Tên ICU được gõ theo nhiều cách (icu a, ICU-A, ICUA, dấu cách thừa ở cuối)	14 dòng	Chuẩn hóa thành ICU A, ICU B, ICU C, ICU D	Nếu không, R sẽ thấy 11 ICU "khác nhau"
4	Ngày nhập viện có ba định dạng: 2023-03-14, 14/03/2023, 14-Mar-2023	365 / 164 / 94 dòng	Đọc từng định dạng một cách tường minh và chuyển tất cả về một định dạng ngày (YYYY-MM-DD); 0 ngày bị lỗi	Định dạng lẫn lộn sẽ bị đọc sai hoặc không đọc được; 03/04 có thể là tháng 3 hoặc tháng 4
5	Giới được mã hóa theo 9 cách (Male, M, m, male, Female, F, f, female, Female)	200 dòng không phải Male/Female	Mã hóa lại thành Female / Male	Một biến, một cách mã hóa
6	Các biến Có/Không lẫn lộn Yes, No, Y, N, yes, no, 1, 0 (ICU B gửi 1/0)	8 biến; khoảng 1900 ô không phải Yes/No	Mã hóa lại thành No / Yes; kiểm tra không mất giá trị nào	1 và Yes phải có cùng một nghĩa

Bước	Vấn đề phát hiện	Số lượng	Cách xử lý	Lý do
7	Mọi cột đều được đọc như văn bản	tất cả	Chuyển số về dạng số (số nguyên cho các giá trị nguyên)	Cần cho việc tính toán
8	Mã giá trị thiếu được dùng như thể là số: lactat 999 và -99, APACHE II -99, MAP 999	Lactat 21 (12 x 999, 9 x -99), APACHE II 6, MAP 4	Đặt là thiếu (NA)	Một giá trị lactat 999 sẽ làm hồng trung bình; thiếu là thiếu
9	Đơn vị: ICU C báo cáo lactat theo mg/dL , các ICU khác theo mmol/L (trung vị ICU C là 23.8 so với khoảng 2.5 ở nơi khác)	126 giá trị (mọi giá trị lactat của ICU C)	Chia các giá trị của ICU C cho 9.01 (1 mmol/L = 9.01 mg/dL), làm tròn 1 chữ số thập phân; trung vị ICU C trở thành 2.65	Một biến, một đơn vị; phát hiện nhờ so sánh trung vị giữa các ICU
10	Giá trị không thể có (nằm ngoài khoảng cho phép trong từ điển dữ liệu)	3: FLU-529 tuổi 210; FLU-277 cân nặng 7 kg; FLU-282 dịch ra ngày 2 là -2990 ml	Đặt là thiếu và gửi query cho đơn vị. Người bệnh vẫn ở lại trong nghiên cứu	Chúng ta không thể đoán giá trị thật (21? 71?); loại bỏ người bệnh sẽ làm sai lệch mẫu
10b	Kiểm tra khoảng giá trị cho mọi biến khác (SOFA, APACHE II, lactat, MAP, giai đoạn AKI, VFD, thời gian nằm ICU, thời gian theo dõi; người sống được theo dõi đến ngày 28)	0 vấn đề	Không cần xử lý	Ghi lại để chứng minh các bước kiểm tra đã được làm
11	Nguồn nhiễm khuẩn ở dạng văn bản tự do	5 nhóm	Cố định thứ tự các nhóm (Respiratory, Urinary, Abdominal, Skin/soft tissue, Other)	Cùng một thứ tự trong mọi bảng
12	Cần các biến dẫn xuất (không có trong tệp thô)	3 biến mới	Tính <code>cum_fluid_balance_d3_ml_kg</code> = (dịch vào ngày 1-3 - dịch ra ngày 1-3) / cân nặng; <code>positive_balance_d3</code> (> 0 ml/kg); <code>aki_any</code> (KDIGO giai đoạn >= 1)	Giá trị dẫn xuất do script tính, không bao giờ gõ tay

Dữ liệu thiếu sau khi làm sạch (giá trị thiếu thật, được báo cáo trong bài báo)

Biến	Số thiếu	Lý do
<code>lactate_mmol_l</code>	93 (15.0%)	Không đo (thường gặp hơn ở người bệnh ít nặng và ở ICU D)
<code>apache_ii</code>	6 (1.0%)	Không ghi nhận
<code>map_mmhg</code>	4 (0.6%)	Không ghi nhận
<code>age</code>	1	Giá trị không thể có được đặt là thiếu (query đang chờ)

Biến	Số thiếu	Lý do
weight_kg	1	Giá trị không thể có được đặt là thiếu (query đang chờ)
fluid_out_d2_ml	1	Giá trị không thể có được đặt là thiếu (query đang chờ)
cum_fluid_balance_d3_ml_kg, positive_balance_d3	2	Hệ quả của cân nặng và dịch ra bị thiếu ở trên

Các query gửi cho đơn vị (mô phỏng)

Query	Người bệnh	Biến	Giá trị trong tệp thô	Tình trạng
Q-001	FLU-529 (ICU B)	Tuổi	210	Đang mở; đặt là thiếu
Q-002	FLU-277 (ICU C)	Cân nặng	7 kg	Đang mở; đặt là thiếu
Q-003	FLU-282 (ICU A)	Dịch ra, ngày 2	-2990 ml	Đang mở; đặt là thiếu
Q-004	ICU C	Đơn vị lactat	mg/dL	Phòng xét nghiệm của đơn vị đã xác nhận; đã quy đổi

Ghi chú giảng dạy: trong một nghiên cứu thật, đơn vị sẽ sửa hồ sơ gốc và cơ sở dữ liệu, rồi script làm sạch được chạy lại. Ở đây các query được để mở để tệp sạch cố định cho khóa học.