

Cấp độ 2: Tự phân tích

Mô hình thống kê cho dữ liệu lâm sàng bằng R

Đơn vị tổ chức: Neudata · #ClearDataClearImpact

> Tình trạng: đề cương. Tài liệu sẽ được xây dựng sau Cấp độ 1.

1. Vì sao có Cấp độ 2

Cấp độ 1 trình bày toàn bộ hành trình nghiên cứu. Giảng viên chạy các phân tích, còn học viên đọc và diễn giải kết quả. Cấp độ 2 dành cho học viên muốn tự chạy các phân tích.

Cấp độ 2 là một khóa thống kê dạy bằng R. Khóa học bao gồm các mô hình hồi quy mà nhà nghiên cứu lâm sàng gặp nhiều nhất, mỗi buổi một nhóm mô hình: hồi quy tuyến tính cho kết cục liên tục, hồi quy logistic cho kết cục nhị phân, mô hình Poisson, nhị thức âm và mô hình thừa số 0 (zero-inflated) cho biến đếm, Kaplan–Meier và mô hình Cox cho thời gian đến biến cố. Với mỗi mô hình, chúng ta đặt cùng một loạt câu hỏi:

- Mô hình này dùng cho những câu hỏi và kết cục nào?
- Cần xem gì trong dữ liệu trước khi xây dựng mô hình (phân tích khám phá dữ liệu, exploratory data analysis, EDA)?
- Các tham số có ý nghĩa gì, theo đơn vị của dữ liệu?
- Mô hình giả định điều gì, và kiểm tra giả định đó như thế nào?
- Làm gì khi một giả định không được đáp ứng?
- Diễn giải và báo cáo kết quả như thế nào?

Kỹ năng thực hành chỉ được dạy ở mức phân tích cần đến: nhập và làm sạch một bộ dữ liệu, mô tả dữ liệu bằng Bảng 1, xử lý dữ liệu thiếu ở mức đơn giản, và viết script R rõ ràng. Xây dựng dự án có thể lặp lại hoàn chỉnh (cấu trúc thư mục, quản lý phiên bản, báo cáo Quarto) và tạo biểu đồ đạt yêu cầu về tệp của tạp chí là chủ đề của Cấp độ 3.

Khi kết thúc Cấp độ 2, học viên có thể:

1. Nhập, làm sạch và mô tả một bộ dữ liệu lâm sàng bằng script R, và lập Bảng 1.
2. Mô tả dữ liệu thiếu và giải thích sự khác nhau giữa phân tích ca đầy đủ (complete-case) và quy gán đa bội.
3. Chọn mô hình hồi quy phù hợp dựa trên loại kết cục và câu hỏi nghiên cứu.
4. Khám phá dữ liệu trước khi xây dựng mô hình, bằng các số liệu tóm tắt và biểu đồ phù hợp với vấn đề phân tích.
5. Xây dựng, kiểm tra, diễn giải và báo cáo mô hình tuyến tính, logistic, mô hình cho biến đếm (Poisson, nhị thức âm, thừa số 0) và mô hình Cox, bao gồm cả nhiễu và tương tác.
6. Phân tích một câu hỏi mới bằng script R của chính mình và viết phần Kết quả từ kết quả phân tích.

2. Đối tượng và điều kiện tham gia

Học viên	8 đến 12 nhân viên y tế (bác sĩ, điều dưỡng, v.v.) của khoa đặt hàng, muốn tự phân tích dữ liệu của mình.
Điều kiện tham gia	(1) Đã hoàn thành Cấp độ 1 . (2) Khóa miễn phí Phân tích Dữ liệu Lâm sàng bằng R của Neudata, hoặc kiến thức cơ bản tương đương: RStudio, đối tượng và data frame, nhập tệp CSV, các hàm tidyverse cơ bản, một biểu đồ đơn giản. Một bài kiểm tra mức sẵn sàng ngắn (30 phút, tự chấm) được gửi trước khóa học.
Thiết bị	Máy tính xách tay cá nhân đã cài R, RStudio và các gói của khóa học trước Buổi 1 (có hướng dẫn cài đặt và một buổi hỗ trợ cài đặt trực tuyến 30 phút). Nên có màn hình thứ hai: một màn hình xem màn hình chia sẻ, một màn hình cho RStudio.
Thời gian cam kết	3 giờ mỗi buổi, cộng khoảng 2 đến 3 giờ bài tập về nhà mỗi tuần; cần sự đồng ý của cấp quản lý trực tiếp để có thời gian dành riêng cho khóa học.

3. Hình thức

Cách tổ chức	Trực tuyến, trực tiếp. Học viên tham gia từ nhà và viết code trên máy tính của mình trong khi giảng viên chia sẻ màn hình.
Số buổi	8 buổi, mỗi buổi 3 giờ , trong khoảng 4 đến 6 tuần (ví dụ hai buổi một tuần, hoặc một buổi một tuần với một tuần nghỉ giữa khóa). Số buổi và thời lượng giữ nguyên như đề cương đầu tiên; nội dung được sắp xếp lại xoay quanh các mô hình thống kê.
Nhiệm vụ mỗi buổi	20 đến 30 phút khái niệm. Khoảng 2 giờ viết code trực tiếp và bài tập: giảng viên viết code, học viên làm theo, rồi tự thực hành. 20 phút tóm tắt, hỏi đáp và giao bài tập về nhà. Một lần nghỉ 15 phút.
Học cùng nhau	Cả lớp học chung như một nhóm, không có phòng thảo luận nhóm nhỏ. Bài tập được làm cá nhân trên máy tính của từng học viên. Học viên gửi câu trả lời và kết quả vào khung chat, và giảng viên (hoặc một học viên tình nguyện chia sẻ màn hình) cùng cả lớp xem lại. Lỗi code được hỗ trợ qua khung chat hoặc khi học viên chia sẻ màn hình.
Cấu trúc mỗi buổi về mô hình	(1) Câu hỏi và loại kết cục. (2) EDA cho vấn đề này : các số liệu tóm tắt và biểu đồ cần xem trước khi dùng mô hình này. (3) Mô hình và ý nghĩa các tham số. (4) Giả định và chẩn đoán mô hình, và cách xử lý khi giả định không được đáp ứng. (5) Diễn giải và bảng báo cáo, kèm một câu viết thành lời. Mỗi buổi đều có code ggplot2 đơn giản cho biểu đồ EDA và biểu đồ kết quả; không có buổi riêng về biểu đồ.
Bộ dữ liệu dùng chung	FLUID-ICU (Course/Data/): bộ dữ liệu thuần tập mô phỏng về nhiễm khuẩn huyết tại ICU (dữ liệu thô và dữ liệu sạch, từ điển dữ liệu, nhật ký làm sạch dữ liệu, kế hoạch phân tích thống kê, script R) đã dùng ở Cấp độ 1. Kết cục dạng biến đếm cần một biến đếm mà FLUID-ICU hiện chưa có (xem Mục 8, các vấn đề còn mở).

Dữ liệu của học viên	Mỗi học viên có thể mang theo một bộ dữ liệu đã loại bỏ thông tin định danh (có phê duyệt đạo đức hoặc bảo vệ dữ liệu cần thiết). Hai buổi gặp riêng trực tuyến, mỗi buổi 30 đến 45 phút (sau Buổi 2 và Buổi 5), giúp điều chỉnh script của khóa học cho dữ liệu của học viên.
Hỗ trợ	Thư mục chung của khóa học (dữ liệu, script khởi đầu, lời giải sau mỗi buổi), một kênh hỏi đáp trực tuyến, và một giờ giải đáp của giảng viên mỗi tuần.
Các gói R chính	tidyverse (readr, dplyr, tidyr, ggplot2), readxl, janitor, gtsummary, flextable, naniar, mice, broom, car, performance, DHARMA, MASS, pscl, pROC, splines, survival, ggsurvfit.

4. Đề cương từng buổi

Buổi 1: Chuẩn bị và mô tả dữ liệu: làm sạch, Bảng 1 và dữ liệu thiếu

- **Khái niệm:** không bao giờ sửa dữ liệu thô; nhật ký làm sạch dữ liệu và từ điển dữ liệu; dữ liệu gọn (tidy data); chọn đúng số liệu tóm tắt (trung bình và SD, trung vị và IQR, n và %); Bảng 1 dùng để làm gì, và vì sao giá trị p không nên có trong Bảng 1 của một RCT.
- **Làm việc với script R:** mỗi bước một script, đánh số (01_clean.R, 02_describe.R); phần đầu script ngắn gọn và chú thích; nạp các gói ở đầu script; đọc dữ liệu từ thư mục của khóa học. Thư mục khóa học được cung cấp sẵn.
- **Làm sạch (tập thô FLUID-ICU):** nhập tệp CSV và Excel; kiểm tra cấu trúc và kiểu dữ liệu; tìm bản ghi trùng lặp; mã hóa lại các mã giá trị thiếu (999, -99); thống nhất các nhóm phân loại (M, Male, 1); chuyển đổi ngày tháng; kiểm tra khoảng giá trị và tính nhất quán; tính cân bằng dịch tích lũy theo ml/kg; đặt các mức của biến factor và nhóm tham chiếu. Mỗi thay đổi đều được ghi vào nhật ký làm sạch dữ liệu.
- **Mô tả:** phân phối và độ lệch (biểu đồ histogram, biểu đồ hộp); Bảng 1 theo nhóm cân bằng dịch bằng **gtsummary** (nhãn, đơn vị, số liệu thống kê phù hợp với từng loại biến, số giá trị thiếu, chú thích); xuất sang Word bằng **flextable**.
- **Dữ liệu thiếu, kiến thức cơ bản:** đếm và trực quan hóa giá trị thiếu (**naniar**); so sánh người bệnh có và không có lactate; MCAR, MAR và MNAR, mỗi loại một slide; phân tích ca đầy đủ giả định điều gì; quy gán đa bội làm gì, trình bày một lần bằng kết quả từ `Course/Data/R/05_missing.R` (lactate: OR phân tích ca đầy đủ 1.39 so với OR sau quy gán 1.45). Học viên không tự chạy quy gán ở Cấp độ 2. Phân tích độ nhạy được nhắc đến như bước tiếp theo và được dạy ở Cấp độ 3.
- **FLUID-ICU:** tệp sạch khớp với tệp sạch của khóa học; Bảng 1 khớp với bộ kết quả của Cấp độ 1.
- **Sản phẩm:** 01_clean.R, 02_describe.R, một nhật ký làm sạch dữ liệu và Bảng 1 dạng bảng Word.

Buổi 2: Hồi quy tuyến tính chuyên sâu

- **Bản đồ mở đầu:** mô hình nào cho kết cục nào (liên tục: tuyến tính; nhị phân: logistic; biến đếm: Poisson hoặc nhị thức âm; thời gian đến biến cố: Cox). Bản đồ này được dùng lại trong mọi buổi.

- **EDA cho kết cục liên tục:** histogram và biểu đồ Q-Q của kết cục; biểu đồ hộp của kết cục theo từng biến dự báo phân loại; biểu đồ phân tán có đường làm trơn cho từng biến dự báo liên tục; kiểm tra độ lệch, giá trị ngoại lai và hiệu ứng sàn hoặc trần.
- **Mô hình:** mô hình như một đường thẳng; ý nghĩa của hệ số chặn và hệ số góc theo đơn vị của dữ liệu; biến dự báo liên tục và phân loại; định tâm (centring); hệ số đã hiệu chỉnh; khoảng tin cậy và giá trị p; R^2 .
- **Nhiều:** chọn biến hiệu chỉnh dựa trên câu hỏi và một sơ đồ nhân quả đơn giản (DAG), không dựa trên giá trị p; ước lượng thô và ước lượng đã hiệu chỉnh.
- **Giả định và chẩn đoán:** tuyến tính, độc lập, phương sai không đổi, phần dư phân phối chuẩn; biểu đồ phần dư và biểu đồ Q-Q (`performance::check_model`); các điểm có ảnh hưởng lớn; đa cộng tuyến (VIF).
- **Khi giả định không được đáp ứng:** chuyển đổi log cho kết cục bị lệch và chuyển ngược lại thành tỷ số trung bình nhân; số hạng phi tuyến bằng spline; khi nào một phép so sánh phi tham số là đủ.
- **Tương tác:** đưa vào, diễn giải và vẽ biểu đồ cho một số hạng tương tác.
- **Báo cáo:** bảng hồi quy bằng `gtsummary`; câu "mỗi 10 ml/kg cân bằng dịch cao hơn có liên quan với ...".
- **FLUID-ICU:** thời gian nằm ICU ở người sống sót (Q5: tỷ số trung bình nhân đã hiệu chỉnh 1.24); số ngày không thở máy (Q3) là ví dụ về một kết cục mà mô hình tuyến tính khớp kém (nhiều giá trị 0 do tử vong, nhiều giá trị 28), sẽ được quay lại ở Buổi 5.
- **Sản phẩm:** `03_linear.R`, một biểu đồ chẩn đoán và một bảng báo cáo kèm phần diễn giải.
Buổi gặp riêng thứ 1 diễn ra sau buổi này.

Buổi 3: Hồi quy logistic chuyên sâu

- **EDA cho kết cục nhị phân:** tỷ lệ kết cục theo từng biến dự báo phân loại (kèm KTC 95%); biểu đồ hộp của các biến dự báo liên tục theo kết cục; tỷ lệ có kết cục theo các nhóm (thập phân vị) của một biến dự báo liên tục, để xem mỗi liên quan có tuyến tính trên thang log-odds hay không; kiểm tra mỗi ô đều có đủ biến cố.
- **Mô hình:** số chênh và tỷ số chênh; hàm liên kết logit; ý nghĩa các hệ số; OR cho biến dự báo liên tục theo một đơn vị có ý nghĩa; OR thô và OR đã hiệu chỉnh (thấy rõ nhiều do chỉ định).
- **Xây dựng mô hình:** số biến cố trên mỗi biến và quá khớp; vì sao chọn biến từng bước (stepwise) gây hiểu sai; nhiều và tương tác (thay đổi tác động), cách kiểm định và báo cáo một tương tác.
- **Giả định và kiểm tra:** tính tuyến tính trên thang logit (so sánh spline với chia nhóm); hiện tượng tách biệt (separation) và dữ liệu thừa; các quan sát có ảnh hưởng lớn; phần dư với **DHARMA**.
- **Cơ bản về hiệu năng mô hình:** khả năng phân biệt (đường cong ROC, AUC với **pROC**) và hiệu chuẩn (quan sát so với dự báo theo nhóm nguy cơ); vì sao những chỉ số này quan trọng với mục đích dự báo hơn là giải thích.
- **Giới thiệu kết cục thứ bậc:** hồi quy logistic số chênh tỷ lệ (proportional odds) cho kết cục có thứ bậc như giai đoạn AKI 0 đến 3 (`MASS::polr`), và giả định chính của mô hình này.
- **Báo cáo:** bảng OR thô và OR đã hiệu chỉnh kèm KTC 95%; biểu đồ rừng (forest plot) của các OR; cách viết theo STROBE.
- **FLUID-ICU:** cân bằng dịch dương và tử vong 28 ngày (Q1: OR thô 2.49, OR đã hiệu chỉnh 1.62), cùng tương tác với suy tim; cân bằng dịch và AKI (Q2).

- **Sản phẩm:** 04_logistic.R, bảng OR thô và OR đã hiệu chỉnh, một biểu đồ rừng và một đoạn diễn giải ngắn.

Buổi 4: Kết cục dạng biến đếm (I): hồi quy Poisson và nhị thức âm

- **Kết cục dạng biến đếm trong nghiên cứu lâm sàng:** số lần nhiễm khuẩn, số lần tái nhập viện, số lần ngã, số ngày hỗ trợ cơ quan, số đợt cấp. Số đếm và tỷ suất; vì sao mô hình tuyến tính và logistic là công cụ không phù hợp.
- **EDA cho kết cục dạng biến đếm:** biểu đồ cột phân phối số đếm (bao nhiêu giá trị 0, đuôi dài đến đâu); trung bình so với phương sai, chung và theo nhóm (kiểm tra ban đầu về quá phân tán); số đếm theo thời gian theo dõi; tỷ suất biến cố theo nhóm.
- **Hồi quy Poisson:** hàm liên kết log; tỷ số tỷ suất (rate ratio) và cách diễn giải; offset cho thời gian theo dõi khác nhau (ví dụ số ngày nằm ICU), và khi nào cần offset.
- **Giả định và kiểm tra:** trung bình bằng phương sai; tính độc lập; kiểm định quá phân tán (overdispersion) và kiểm tra phần dư (**DHARMA**, `performance::check_overdispersion`); quá phân tán ảnh hưởng thế nào đến sai số chuẩn và giá trị p.
- **Khi mô hình Poisson không phù hợp:** hồi quy nhị thức âm (`MASS::glm.nb`); sai số chuẩn vững (sandwich) là một cách khác; so sánh mô hình Poisson và nhị thức âm (AIC, biểu đồ phần dư, số đếm quan sát so với dự báo).
- **Báo cáo:** bảng tỷ số tỷ suất kèm KTC 95%; câu "tỷ suất ... cao gấp X lần ...".
- **FLUID-ICU:** một kết cục dạng biến đếm mô phỏng (đề xuất: số lần nhiễm khuẩn mắc phải tại ICU, với thời gian nằm ICU làm offset; xem các vấn đề còn mở).
- **Sản phẩm:** 05_count.R, so sánh mô hình Poisson và nhị thức âm, và bảng tỷ số tỷ suất kèm phần diễn giải.

Buổi 5: Kết cục dạng biến đếm (II): thừa giá trị 0 và mô hình thừa số 0

- **Giá trị 0 dư thừa đến từ đâu:** hai loại giá trị 0 (người bệnh không bao giờ có thể có biến cố, và người bệnh có thể có nhưng đã không có); vì sao khi đó mô hình Poisson hoặc nhị thức âm ước tính thấp số giá trị 0.
- **EDA cho trường hợp thừa giá trị 0:** số giá trị 0 quan sát được so với số kỳ vọng theo mô hình Poisson hoặc nhị thức âm; phân phối số đếm tách theo một nhóm có khả năng là "giá trị 0 cấu trúc"; biểu đồ rootogram so sánh số đếm quan sát và số đếm từ mô hình.
- **Mô hình thừa số 0:** Poisson thừa số 0 (ZIP) và nhị thức âm thừa số 0 (ZINB) với **pscl** (`zeroinfl`); hai phần của mô hình (xác suất là một giá trị 0 cấu trúc, và số đếm ở những người còn lại) và cách diễn giải từng phần.
- **Mô hình hurdle:** khác với mô hình thừa số 0 như thế nào, và cách chọn giữa hai loại dựa trên ý nghĩa lâm sàng của giá trị 0.
- **Chọn và kiểm tra mô hình:** so sánh các mô hình Poisson, nhị thức âm, ZIP, ZINB và hurdle (AIC, số giá trị 0 quan sát so với dự báo, rootogram); vì sao nên chọn mô hình trước hết dựa trên cơ chế tạo ra dữ liệu, sau đó mới đến các chỉ số khớp mô hình.
- **Biến đếm có giới hạn:** số ngày không thở máy là số ngày trên tổng 28 ngày, với giá trị 0 cho người tử vong. Chúng ta thảo luận vì sao biến này khó mô hình hóa (giá trị 0, mức trần 28, tử vong là biến cố cạnh tranh) và các lựa chọn trong y văn: mô hình hurdle hoặc thừa số 0, hoặc mô hình nhị thức thừa số 0 (hay beta-nhị thức) cho "số ngày trên 28 ngày". Mô hình nhị thức thừa số 0 chỉ được trình diễn, không thực hành.
- **Báo cáo:** cần báo cáo gì cho một mô hình hai phần; diễn giải bằng ngôn ngữ dễ hiểu.

- **FLUID-ICU:** phân tích lại số ngày không thở máy (Q3) bằng mô hình hurdle và so sánh với mô hình tuyến tính ở Buổi 2; một biến đếm mô phỏng có nhiều giá trị 0 (đề xuất: số ngày điều trị thay thế thận (RRT) tính đến ngày 28; xem các vấn đề còn mở).
- **Sản phẩm:** 06_zero_inflated.R và một bản so sánh ngắn các mô hình kèm một đoạn diễn giải.

Buổi 6: Phân tích sống còn: Kaplan–Meier và mô hình Cox

- **Dữ liệu thời gian đến biến cố:** dữ liệu bị kiểm duyệt (censoring), và vì sao bỏ qua nó sẽ cho kết quả sai; hàm sống còn và hàm nguy cơ tức thời bằng lời đơn giản; cách sắp xếp dữ liệu để phân tích (thời gian và tình trạng biến cố).
- **EDA cho dữ liệu thời gian đến biến cố:** thời gian theo dõi và số biến cố theo nhóm; kiểm duyệt xảy ra khi nào và có khác nhau giữa các nhóm hay không; đường cong Kaplan–Meier theo từng biến dự báo phân loại chính, kèm số người còn nguy cơ (**ggsurvfit**).
- **Kaplan–Meier và kiểm định log-rank:** tỷ lệ sống còn tại một thời điểm cố định kèm KTC 95%; thời gian sống còn trung vị; kiểm định log-rank so sánh điều gì.
- **Mô hình Cox:** tỷ số nguy cơ tức thời và ý nghĩa của nó (cũng như điều nó không có nghĩa); HR thô và HR đã hiệu chỉnh; biến dự báo liên tục theo một đơn vị có ý nghĩa.
- **Giả định nguy cơ tỷ lệ:** ý nghĩa của giả định; kiểm tra bằng phần dư Schoenfeld (`cox.zph`) và biểu đồ log-minus-log; xử lý khi giả định không được đáp ứng (phân tầng; giới thiệu cách chia thời gian theo dõi). Biến số thay đổi theo thời gian và nguy cơ cạnh tranh được nhắc đến và được dạy ở Cấp độ 3 (Học phần D).
- **Các kiểm tra khác:** dạng hàm của biến dự báo liên tục; các quan sát có ảnh hưởng lớn.
- **Báo cáo:** biểu đồ Kaplan–Meier kèm số người còn nguy cơ; bảng HR thô và HR đã hiệu chỉnh kèm KTC 95%; nêu việc kiểm tra giả định nguy cơ tỷ lệ trong phần Phương pháp.
- **FLUID-ICU:** cân bằng dịch dương và sống còn 28 ngày (Q1: Kaplan–Meier 81.6% so với 64.1% tại ngày 28; HR đã hiệu chỉnh 1.46; kiểm định tổng thể giả định nguy cơ tỷ lệ $p = 0.10$), so sánh với kết quả hồi quy logistic ở Buổi 3.
- **Sản phẩm:** 07_survival.R, một biểu đồ Kaplan–Meier, bảng HR kèm kiểm tra giả định nguy cơ tỷ lệ và một đoạn diễn giải ngắn. *Buổi gặp riêng thứ 2 diễn ra sau buổi này.*

Buổi 7: Dự án phân tích (I): chọn và chạy đúng mô hình

- **Chuẩn bị:** mỗi học viên nhận một câu hỏi phân tích khác nhau kèm một bản mô tả ngắn (câu hỏi, kết cục, bộ dữ liệu, nhóm mô hình dự kiến). Các câu hỏi trải đều trên các loại mô hình đã học: tuyến tính, logistic, biến đếm (Poisson, nhị thức âm hoặc thừa số 0) và Cox. Dữ liệu là FLUID-ICU, một bộ dữ liệu mô phỏng thứ hai chuẩn bị cho khóa học, hoặc dữ liệu của chính học viên nếu giảng viên đồng ý.
- **Trên lớp:** giảng viên cùng cả lớp làm một câu hỏi ngắn từ đầu đến cuối (câu hỏi, EDA, chọn mô hình, kiểm tra, diễn giải) làm mẫu cho quy trình. Sau đó học viên làm cá nhân câu hỏi của mình bằng script R, giảng viên trả lời câu hỏi qua khung chat và chia sẻ màn hình. Giữa buổi, một vài học viên chia sẻ màn hình để trình bày EDA và lựa chọn mô hình của mình với cả lớp.
- **Bảng kiểm quy trình:** làm sạch và mô tả dữ liệu; Bảng 1; EDA phù hợp với kết cục; chọn mô hình và các biến hiệu chỉnh; kiểm tra giả định; diễn giải; lưu bảng và biểu đồ.
- **Sản phẩm (bài tập về nhà):** một bộ script R chạy từ dữ liệu đến tất cả bảng và biểu đồ, và một ghi chú ngắn về mô hình đã chọn và các kiểm tra đã làm.

Buổi 8: Dự án phân tích (II): viết phần Kết quả và chia sẻ

- **Viết phần Kết quả:** từ kết quả phân tích đến văn bản; nêu độ lớn tác động kèm KTC 95% trước giá trị p; ngôn ngữ chừng mực; các con số khớp nhau giữa văn bản, bảng và biểu đồ; các mục STROBE (hoặc CONSORT) cho phần Kết quả; đoạn phân tích thống kê cho phần Phương pháp.
- **Trên lớp:** giảng viên cùng cả lớp viết đoạn Kết quả cho câu hỏi mẫu. Sau đó học viên tự viết bản nháp phần Kết quả của mình.
- **Chia sẻ lại:** trong giờ cuối, một số học viên (nếu có thể thì mỗi loại mô hình một người) trình bày trước lớp trong 5 phút câu hỏi, lựa chọn mô hình, các kiểm tra và kết quả chính của mình. Giảng viên và cả lớp nhận xét về lựa chọn mô hình, các kiểm tra và cách diễn đạt.
- **Sản phẩm:** một phần Kết quả (khoảng 300 đến 500 từ) kèm Bảng 1, một bảng kết quả chính và một biểu đồ, cùng các script R đã tạo ra chúng.

5. Sản phẩm từng buổi (tóm tắt)

Buổi	Sản phẩm
1	Script làm sạch và mô tả dữ liệu, nhật ký làm sạch dữ liệu, Bảng 1 (Word)
2	Script mô hình tuyến tính, biểu đồ chẩn đoán, bảng báo cáo và phần diễn giải
3	Script mô hình logistic, bảng OR thô và OR đã hiệu chỉnh, biểu đồ rừng, phần diễn giải
4	Script mô hình biến đếm, so sánh Poisson và nhị thức âm, bảng tỷ số tỷ suất
5	Script mô hình thừa số 0 hoặc hurdle và so sánh mô hình kèm phần diễn giải
6	Script phân tích sống còn, biểu đồ Kaplan–Meier, bảng HR kèm kiểm tra giả định nguy cơ tỷ lệ, phần diễn giải
7	Script R cho câu hỏi phân tích cá nhân và ghi chú về lựa chọn mô hình và các kiểm tra
8	Phần Kết quả kèm Bảng 1, một bảng kết quả và một biểu đồ, cùng các script

Mỗi sản phẩm được nộp vào thư mục khóa học trước buổi tiếp theo; giảng viên phản hồi ngắn bằng văn bản.

6. Dự án cuối khóa (không bắt buộc)

Trong vòng **4 tuần sau Buổi 8**, học viên muốn đi xa hơn có thể nộp một phân tích trên **dữ liệu của chính mình** (ưu tiên, khi đã được phê duyệt) hoặc trên một câu hỏi thứ hai của khóa học, với loại mô hình khác với câu hỏi ở Buổi 7 và 8:

1. **Script R** chạy từ dữ liệu sạch đến tất cả bảng và biểu đồ mà không lỗi, kèm một kế hoạch phân tích ngắn viết trước phân tích chính.
2. Một **bản nháp phần Kết quả** (khoảng 500 đến 800 từ) kèm Bảng 1, một bảng kết quả chính và một biểu đồ, cùng đoạn phân tích thống kê cho phần Phương pháp.

Không yêu cầu một dự án có thể lặp lại hoàn chỉnh với báo cáo Quarto; nội dung đó được dạy ở Cấp độ 3.

7. Đánh giá

Thành phần	Tỷ trọng	Cách đánh giá
Sản phẩm từng buổi (Buổi 1 đến 6)	40%	Hoàn thành và chạy được (đạt hoặc không đạt cho từng sản phẩm, kèm phản hồi ngắn)
Dự án phân tích: script và mô hình (Buổi 7)	30%	Script chạy được; mô hình phù hợp với câu hỏi và kết cục; đã làm EDA; đã kiểm tra giả định; diễn giải đúng
Dự án phân tích: phần Kết quả (Buổi 8)	30%	Độ lớn tác động kèm KTC 95%; bảng và văn bản khớp nhau; đủ các mục của hướng dẫn báo cáo; ngôn ngữ chừng mực

Điểm số dùng để phản hồi, không dùng để xếp hạng.

8. Tài liệu cần xây dựng và các vấn đề còn mở

Tài liệu	Nội dung
Slide (EN + VN)	8 bộ slide khái niệm ngắn (mỗi bộ 20 đến 30 slide) kèm ghi chú cho người trình bày; mỗi bộ về mô hình theo cùng năm phần (câu hỏi, EDA, mô hình, giả định, báo cáo)
Script R	Script khởi đầu (có chỗ trống) và script hoàn chỉnh của giảng viên cho từng buổi, xây dựng từ các script FLUID-ICU trong <code>Course/Data/R/</code> ; code ggplot2 làm sẵn cho biểu đồ EDA và biểu đồ kết quả của từng mô hình
Bài tập	Mỗi buổi một phiếu bài tập (phần cơ bản và phần nâng cao) trên FLUID-ICU, kèm một bài "dữ liệu của bạn"
Lời giải	Lời giải đầy đủ kèm kết quả, phát sau mỗi buổi
Mô tả dự án phân tích	8 đến 12 câu hỏi cá nhân trải trên các loại mô hình (tuyến tính, logistic, biến đếm, thừa số 0, Cox), mỗi câu kèm dữ liệu và hướng dẫn chấm điểm
Bảng kiểm	Bản đồ chọn mô hình; bảng kiểm giả định cho từng mô hình; bảng kiểm phần Kết quả (STROBE hoặc CONSORT)
Hướng dẫn chuẩn bị	Hướng dẫn cài đặt, danh sách gói và bài kiểm tra mức sẵn sàng

Các vấn đề còn mở

- **FLUID-ICU cần biến đếm.** Bộ dữ liệu chưa có kết cục dạng biến đếm thực sự. Đề xuất bổ sung vào chương trình tạo dữ liệu (`Course/Data/R/00_generate_fluid_icu.R`), từ điển dữ liệu và `key_results.md`: (1) `icu_infections_n`, số lần nhiễm khuẩn mắc phải tại ICU trong thời gian nằm ICU, có quá phân tán, với `icu_los_days` làm offset (Buổi 4); (2) `rrt_days_28`, số ngày điều trị thay thế thận tính đến ngày 28, bằng 0 với người bệnh không điều trị thay thế thận (Buổi 5; liên kết với biến `rrt` hiện có). Các phương án khác: số lần tái nhập ICU trong 90 ngày, hoặc số ngày hỗ trợ cơ quan. Trước khi có các biến này, Buổi 5 có thể dùng số ngày không thở máy, còn Buổi 4 cần một bộ dữ liệu biến đếm mô phỏng nhỏ riêng.
- **Có thể cần một bộ dữ liệu mô phỏng thứ hai** cho các câu hỏi dự án phân tích, để mỗi học viên có một câu hỏi khác nhau và không phải ai cũng dùng FLUID-ICU.