

Cấp độ 3: Chủ đề nâng cao

Phương pháp chuyên sâu cho dữ liệu lâm sàng phức tạp: hội thảo và tư vấn

Đơn vị tổ chức: Neudata · #ClearDataClearImpact

> **Tình trạng:** đề cương. Các học phần được xây dựng theo yêu cầu, sau khi Cấp độ 1 và Cấp độ 2 đã diễn ra.

1. Cấp độ 3 là gì

Dữ liệu lâm sàng thực tế thường phức tạp hơn bộ dữ liệu thuần tập dùng ở Cấp độ 1 và 2: đo lặp lại trên cùng người bệnh, người bệnh theo cụm trong khoa phòng hoặc bệnh viện, các kết cục cạnh tranh, sổ đăng ký thường quy quy mô lớn, hoặc câu hỏi về dự báo và chẩn đoán thay vì mối liên quan. Cấp độ 3 gồm các học phần chuyên sâu, mỗi học phần gắn với loại dữ liệu hoặc câu hỏi mà học viên gặp trong công việc của mình. Cấp độ 3 cũng bao gồm các chủ đề về kỹ năng làm nghiên cứu đã chuyển ra khỏi Cấp độ 2 để Cấp độ 2 tập trung vào mô hình thống kê: dự án có thể lặp lại, quản lý phiên bản, báo cáo Quarto, biểu đồ đạt yêu cầu của tạp chí, dữ liệu thiếu và phân tích độ nhạy chuyên sâu (Học phần G và H). Neudata tổ chức các học phần dưới dạng hội thảo và tư vấn: một đợt giảng dạy ngắn, tập trung, sau đó là thực hành trên chính dự án của học viên, với sự hỗ trợ của Neudata cho đến khi có một kế hoạch phân tích hoặc một phân tích vững chắc.

2. Hình thức

Đối tượng	Học viên đã hoàn thành Cấp độ 2 (hoặc nhà nghiên cứu có kinh nghiệm tương đương về R và hồi quy), thường 3 đến 10 người mỗi hội thảo; nhà nghiên cứu cá nhân hoặc nhóm dự án với hình thức tư vấn.
Điều kiện tham gia	Cấp độ 2 (hoặc tương đương): làm việc với script R, làm sạch dữ liệu, Bảng 1, các mô hình tuyến tính, logistic, biến đếm (Poisson, nhị thức âm, thừa số 0) và Cox cùng giả định và chẩn đoán của chúng; kiến thức cơ bản về dữ liệu thiếu. Rất nên có một câu hỏi hoặc bộ dữ liệu thực tế.
Hình thức hội thảo	1 hoặc 2 buổi trực tuyến, mỗi buổi 3 giờ: khái niệm, rồi một ví dụ mẫu bằng R, rồi học viên áp dụng trên dữ liệu của mình. Cả lớp học chung như một nhóm (không có phòng thảo luận nhóm nhỏ); phần áp dụng làm cá nhân, và học viên chia sẻ kết quả lại với cả lớp.
Hình thức tư vấn	Một buổi họp xác định phạm vi, sau đó 2 đến 4 buổi họp trực tuyến tiếp theo trong 1 đến 3 tháng; học viên tự làm phân tích, Neudata xem xét kế hoạch, code và cách diễn giải.
Phần mềm	R và RStudio; các gói riêng cho từng học phần (nêu tên trong từng học phần). Mọi học phần đều dùng ggplot2 cho biểu đồ EDA và biểu đồ kết quả.

Kết quả đầu ra	Một kế hoạch phân tích thống kê, một script phân tích đã được xem xét và bản nháp phần Phương pháp/Kết quả cho dự án của học viên, hoặc một bản thẩm định có cấu trúc đối với các học phần tập trung vào thẩm định.
Cách đăng ký	Khoa hoặc nhóm dự án liên hệ Neudata với một mô tả ngắn về câu hỏi và dữ liệu; Neudata đề xuất (các) học phần và hình thức.

3. Nguyên tắc chung cho mọi học phần

Các nguyên tắc này áp dụng cho mọi học phần Cấp độ 3 và mọi buổi trong học phần.

- 1. **Mỗi bài giảng đều có phần EDA.** Mỗi buổi có một phần về phân tích khám phá dữ liệu (exploratory data analysis, EDA) cho vấn đề phân tích của buổi đó: cần xem những số liệu tóm tắt và biểu đồ nào trước khi xây dựng mô hình, và chúng có thể cho thấy điều gì. Ví dụ: biểu đồ spaghetti và tình trạng thiếu dữ liệu theo thời gian với số đo lặp lại (Học phần E); kiểm theo dõi và kiểm duyệt trong phân tích sống còn (Học phần D); phân phối của xét nghiệm cần đánh giá theo tình trạng bệnh (Học phần C); đặc điểm các nghiên cứu và độ lớn tác động trước khi gộp (Học phần A).
- 2. **Mỗi bài giảng đều có code vẽ biểu đồ.** Mỗi buổi cung cấp code ggplot2 dùng được ngay cho các biểu đồ mà vấn đề của buổi đó cần, đạt chuẩn tạp chí ngay từ đầu. Ví dụ: biểu đồ rừng và biểu đồ phễu (A), biểu đồ hiệu chuẩn (B), đường cong ROC (C), đường cong tỷ lệ mới mắc tích lũy (D), quỹ đạo từng cá nhân và giá trị dự báo từ mô hình (E). Các quy tắc chung cho biểu đồ đăng tạp chí (tệp vector và raster, kích thước, độ phân giải và phông chữ) được dạy một lần ở Học phần G rồi áp dụng trong mọi học phần.
- 3. **Có thể lặp lại ngay từ đầu.** Bài tập của các học phần được làm trong một dự án có thể lặp lại (Học phần G), để người khác có thể chạy lại phân tích và báo cáo.

4. Các học phần

Học phần A: Tổng quan hệ thống và phân tích gộp

Mục tiêu. Học viên có thể (1) viết và đăng ký (PROSPERO) đề cương tổng quan hệ thống; (2) xây dựng chiến lược tìm kiếm có thể lặp lại trên nhiều cơ sở dữ liệu và sàng lọc tài liệu theo cặp; (3) trích xuất dữ liệu và đánh giá nguy cơ sai lệch (RoB 2, ROBINS-I); (4) gộp các ước lượng tác động bằng mô hình tác động cố định và tác động ngẫu nhiên, diễn giải tính không đồng nhất (I^2 , khoảng dự báo) và vẽ biểu đồ rừng và biểu đồ phễu; và (5) báo cáo theo PRISMA 2020 và xếp hạng mức độ tin cậy của bằng chứng (GRADE).

Điều kiện tham gia. Cấp độ 2; một câu hỏi tổng quan (PICO) đã thống nhất với Neudata. Các gói: meta, metafor.

Buổi	Nội dung	Sản phẩm
A1: Đề cương, tìm kiếm và sàng lọc	Câu hỏi tổng quan và tiêu chí chọn; PROSPERO; chiến lược tìm kiếm trên nhiều cơ sở dữ liệu; sàng lọc kép; sơ đồ PRISMA; biểu mẫu trích xuất dữ liệu	Đề cương đã đăng ký (hoặc sẵn sàng đăng ký) và chiến lược tìm kiếm

Buổi	Nội dung	Sản phẩm
A2: Sai lệch, gộp và báo cáo	Đánh giá nguy cơ sai lệch; gộp OR/RR/hiệu số trung bình bằng R; tác động cố định và tác động ngẫu nhiên; I^2 , phân nhóm và hồi quy gộp (meta-regression, giới thiệu); hiệu ứng nghiên cứu nhỏ và biểu đồ phễu; GRADE; PRISMA 2020	Biểu đồ rừng và bản nháp phần Kết quả cho tổng quan của học viên

Học phần B: Mô hình dự báo và học máy

Mục tiêu. Học viên có thể (1) phân biệt câu hỏi dự báo với câu hỏi nhân quả; (2) tính cỡ mẫu cần thiết để xây dựng mô hình; (3) xây dựng mô hình dự báo logistic hoặc sống còn, đánh giá khả năng phân biệt và hiệu chuẩn, và thẩm định nội bộ (bootstrap); (4) hiểu các phương pháp học máy bổ sung được gì và gây hiểu sai ở đâu (rò rỉ dữ liệu, quá khớp, tính công bằng); và (5) báo cáo theo TRIPOD+AI và thẩm định bằng PROBAST.

Điều kiện tham gia. Cấp độ 2 (Buổi 3, hồi quy logistic chuyên sâu). Các gói: rms, pmsampsize, pROC, glmnet, tidymodels (giới thiệu).

Buổi	Nội dung	Sản phẩm
B1: Xây dựng mô hình đúng cách	Dự báo và nhân quả; các biến dự báo ứng viên; cỡ mẫu cho mô hình dự báo; xây dựng mô hình trên FLUID-ICU hoặc dữ liệu của học viên; biểu đồ khả năng phân biệt và hiệu chuẩn; thẩm định nội bộ bằng bootstrap; trình bày mô hình (nomogram, biểu đồ nguy cơ)	Báo cáo mô hình với AUC đã hiệu chỉnh độ lệch quan và một biểu đồ hiệu chuẩn
B2: Học máy và thẩm định	Hồi quy có phạt (penalised regression) và một phương pháp dựa trên cây quyết định, so sánh với hồi quy logistic; rò rỉ dữ liệu; thẩm định ngoại và cập nhật mô hình; tính công bằng; thẩm định một mô hình AI/học máy đã công bố bằng TRIPOD+AI và PROBAST	Bản thẩm định có cấu trúc một mô hình đã công bố, hoặc kế hoạch thẩm định cho mô hình của học viên

Học phần C: Nghiên cứu độ chính xác chẩn đoán

Mục tiêu. Học viên có thể (1) thiết kế một nghiên cứu độ chính xác chẩn đoán (xét nghiệm cần đánh giá, tiêu chuẩn tham chiếu, phổ bệnh, làm mù, xác minh); (2) tính độ nhạy, độ đặc hiệu, giá trị tiên đoán và tỷ số khả dĩ kèm KTC 95%, và vẽ đường cong ROC kèm AUC; (3) chọn và lý giải một điểm cắt; (4) so sánh hai xét nghiệm trên cùng người bệnh; và (5) báo cáo theo STARD 2015 và thẩm định bằng QUADAS-2.

Điều kiện tham gia. Cấp độ 2. Các gói: epiR, pROC.

Buổi	Nội dung	Sản phẩm
C1: Thiết kế, phân tích và báo cáo	Câu hỏi và thiết kế nghiên cứu chẩn đoán; sai lệch phổ bệnh, sai lệch xác minh và sai lệch do gộp xét nghiệm vào tiêu chuẩn tham chiếu (incorporation bias); bảng độ chính xác kèm KTC; giá trị tiên đoán ở các tỷ lệ hiện mắc khác nhau; ROC/AUC và điểm cắt (ví dụ lactate cho tử vong 28 ngày); so sánh hai xét nghiệm (ROC ghép cặp); STARD 2015 và QUADAS-2	Phân tích độ chính xác hoàn chỉnh và một đề cương nghiên cứu dựa trên STARD
C2 (không bắt buộc): Câu hỏi chẩn đoán của bạn	Đề cương cho một nghiên cứu độ chính xác ở quy mô khoa; cỡ mẫu cho độ nhạy và độ đặc hiệu; biểu mẫu thu thập dữ liệu	Dàn ý đề cương

Học phần D: Phân tích sống còn nâng cao

Mục tiêu. Học viên có thể (1) vượt ra ngoài mô hình Cox của Cấp độ 2 khi giả định của nó không được đáp ứng (nguy cơ không tỷ lệ, tác động thay đổi theo thời gian); (2) dùng biến số thay đổi theo thời gian và tránh sai lệch thời gian bất tử (immortal-time bias); (3) phân tích nguy cơ cạnh tranh (hàm tỷ lệ mới mắc tích lũy, mô hình theo nguyên nhân và mô hình Fine–Gray), ví dụ ra khỏi ICU còn sống so với tử vong; (4) trình bày thời gian sống trung bình giới hạn (restricted mean survival time) như một lựa chọn thay cho HR; và (5) báo cáo rõ ràng các phân tích thời gian đến biến cố.

Điều kiện tham gia. Cấp độ 2, bao gồm Buổi 6 (Kaplan–Meier, mô hình Cox và kiểm tra giả định nguy cơ tỷ lệ). Các gói: survival, survminer, tidycmprsk, cmprsk.

Buổi	Nội dung	Sản phẩm
D1: Vượt ra ngoài mô hình Cox cơ bản	EDA cho dữ liệu thời gian đến biến cố khi phơi nhiễm thay đổi; khi giả định nguy cơ tỷ lệ không được đáp ứng: phân tầng và tác động thay đổi theo thời gian; biến số thay đổi theo thời gian và định dạng quá trình đếm (counting process); sai lệch thời gian bất tử và phân tích mốc (landmark); thời gian sống trung bình giới hạn	Phân tích Cox kèm kiểm tra giả định nguy cơ tỷ lệ và một biến số thay đổi theo thời gian
D2: Nguy cơ cạnh tranh	Vì sao 1 – KM ước tính quá cao nguy cơ khi có biến cố cạnh tranh; tỷ lệ mới mắc tích lũy; mô hình theo nguyên nhân và mô hình Fine–Gray, khi nào dùng mỗi loại; vẽ biểu đồ và báo cáo	Phân tích nguy cơ cạnh tranh trên FLUID-ICU hoặc dữ liệu của học viên

Học phần E: Dữ liệu dọc và mô hình đa tầng (hỗn hợp)

Mục tiêu. Học viên có thể (1) sắp xếp dữ liệu đo lặp lại (dạng dài và dạng rộng) và vẽ quỹ đạo của từng cá nhân; (2) xây dựng mô hình hỗn hợp tuyến tính và logistic với hệ số chặn ngẫu nhiên và hệ số góc ngẫu nhiên; (3) diễn giải tác động cố định, tác động ngẫu nhiên và hệ số tương quan nội lớp; (4) xử lý dữ liệu thiếu và bỏ cuộc trong số đo lặp lại; và (5) biết khi nào phương trình ước lượng tổng quát (GEE) là lựa chọn tốt hơn.

Điều kiện tham gia. Cấp độ 2 (mô hình tuyến tính, logistic và biến đếm). Các gói: lme4, glmmTMB, geepack, ggeffects.

Buổi	Nội dung	Sản phẩm
E1: Số đo lặp lại theo thời gian	Dữ liệu dạng dài và dạng rộng; biểu đồ spaghetti; vì sao hồi quy thông thường không đúng với số đo lặp lại; mô hình hỗn hợp tuyến tính cho số đo hằng ngày (ví dụ cân bằng dịch hoặc SOFA ngày 1 đến 3); hệ số chặn và hệ số góc ngẫu nhiên; thời gian là một biến dự báo	Phân tích mô hình hỗn hợp cho một kết cục đo lặp lại kèm một đoạn Kết quả
E2: Kết cục theo cụm và GEE	Kết cục nhị phân đo lặp lại hoặc theo cụm; mô hình hỗn hợp logistic; GEE, tác động trung bình quần thể và tác động theo cá thể; ICC; kiểm tra mô hình; báo cáo	So sánh ước lượng từ mô hình hỗn hợp và GEE cho một câu hỏi

Học phần F: Dữ liệu phức tạp: thiết kế theo cụm, số đo lặp lại và sổ đăng ký lớn

Mục tiêu. Học viên có thể (1) nhận biết dữ liệu theo cụm (người bệnh trong khoa phòng, ICU hoặc bệnh viện) và hệ quả của nó đối với sai số chuẩn và cỡ mẫu; (2) phân tích dữ liệu thử nghiệm ngẫu nhiên theo cụm và dữ liệu đa trung tâm (tác động của trung tâm, hệ số thiết kế); (3) làm việc với bộ dữ liệu thường quy lớn và sổ đăng ký (liên kết dữ liệu, kiểm tra chất lượng dữ liệu ở quy mô lớn, các hệ thống mã hóa như ICD, xử lý dữ liệu hiệu quả trong R); và (4) lập kế hoạch phân tích nhân quả cho dữ liệu quan sát (DAG, điểm xu hướng, giới thiệu ý tưởng thử nghiệm mục tiêu (target trial)).

Điều kiện tham gia. Cấp độ 2; nên học trước Học phần E. Các gói: data.table hoặc arrow, MatchIt, WeightIt, cobalt, dagitty.

Buổi	Nội dung	Sản phẩm
F1: Dữ liệu theo cụm và đa trung tâm	Hệ số thiết kế và ICC; trung tâm là tác động cố định hay tác động ngẫu nhiên; sai số chuẩn vững theo cụm; phân tích một nghiên cứu ngẫu nhiên theo cụm hoặc đa trung tâm; báo cáo (CONSORT mở rộng cho thử nghiệm theo cụm)	Kế hoạch phân tích cho một bộ dữ liệu theo cụm hoặc đa trung tâm
F2: Sổ đăng ký lớn và dữ liệu thường quy	Liên kết dữ liệu và loại bỏ thông tin định danh; chất lượng dữ liệu ở quy mô lớn; hệ thống mã hóa; xử lý dữ liệu hiệu quả; câu hỏi nhân quả từ dữ liệu thường quy (DAG, điểm xu hướng với ghép cặp, trọng số và kiểm tra cân bằng, và giới thiệu mô phỏng thử nghiệm mục tiêu); báo cáo theo RECORD	Kế hoạch quản lý và phân tích dữ liệu cho một câu hỏi dựa trên sổ đăng ký

Học phần G: Nghiên cứu có thể lặp lại và kết quả đạt chuẩn công bố

Học phần này gồm các chủ đề về khả năng lặp lại và biểu đồ được chuyển từ Cấp độ 2. Học phần này được khuyến nghị học đầu tiên trong Cấp độ 3 cho mọi người, vì các học phần khác dùng cấu trúc dự án và quy tắc biểu đồ của nó.

Mục tiêu. Học viên có thể (1) tổ chức một phân tích thành một dự án R có thể lặp lại mà người khác chạy lại được (phép thử "chạy lại sau một năm"); (2) lưu lịch sử công việc bằng quản lý phiên bản; (3) viết một báo cáo Quarto trong đó mọi con số, bảng và biểu đồ đều do code tạo ra; và (4) tạo biểu đồ đạt yêu cầu của tạp chí.

Điều kiện tham gia. Cấp độ 2. Các gói: here, renv, quarto, gtsummary, flextable, ggplot2, ragg, svglite, patchwork; Git qua RStudio hoặc GitHub Desktop.

Buổi	Nội dung	Sản phẩm
G1: Dự án có thể lặp lại và quản lý phiên bản	Vì sao khả năng lặp lại quan trọng; quy trình từ dữ liệu thô đến dữ liệu sạch, phân tích và báo cáo; một RStudio Project với cấu trúc thư mục chuẩn (data/raw, data/clean, R/, output/, report/); đường dẫn với here; phần đầu script, script đánh số, nạp gói ở đầu script, sessionInfo() và renv; quản lý phiên bản bằng Git (commit, lịch sử, khôi phục, một bản sao từ xa trên GitHub hoặc nền tảng tương tự), với các phiên bản ghi ngày và nhật ký thay đổi là phương án tối thiểu thay thế	Một dự án FLUID-ICU được quản lý phiên bản, chạy lại được từ tệp thô trên một máy tính khác
G2: Báo cáo Quarto và biểu đồ đạt chuẩn tạp chí	Một báo cáo Quarto chạy toàn bộ quy trình và xuất ra Word và HTML; con số chèn trực tiếp trong câu; mỗi con số chỉ có một nguồn; bảng theo phong cách tạp chí; biểu đồ cho tạp chí: tệp vector (PDF, EPS, SVG) và raster (TIFF, PNG) và khi nào cần loại nào, chiều rộng và chiều cao tính theo mm cho một cột và hai cột, độ phân giải (300 đến 600 dpi hoặc cao hơn cho hình nét, theo quy định của tạp chí), phông chữ và cỡ chữ tối thiểu, bảng màu thân thiện với người mù màu, biểu đồ nhiều ô với patchwork, xuất tệp bằng ggsave; đối chiếu biểu đồ với hướng dẫn tác giả của tạp chí	Một báo cáo Quarto đã xuất và hai biểu đồ xuất theo yêu cầu của một tạp chí mục tiêu

Học phần H: Dữ liệu thiếu và phân tích độ nhạy chuyên sâu

Mục tiêu. Học viên có thể (1) xây dựng và kiểm tra một mô hình quy gán đa bội (các biến cần đưa vào, đưa kết cục vào mô hình quy gán, số lần quy gán, sự hội tụ và phân phối của giá trị quy gán, quy tắc Rubin); (2) quy gán cẩn thận các số hạng phi tuyến, tương tác và dữ liệu đa tầng hoặc dữ liệu dọc; (3) lập kế hoạch và thực hiện phân tích độ nhạy cho các giả định về dữ liệu thiếu (bao gồm các kịch bản MNAR với hiệu chỉnh delta), giá trị ngoại lai, định nghĩa phơi nhiễm và kết cục, và nhiễu chưa đo lường (giới thiệu E-value); và (4) báo cáo dữ liệu thiếu và phân tích độ nhạy ngắn gọn, rõ ràng.

Điều kiện tham gia. Cấp độ 2 (kiến thức cơ bản về dữ liệu thiếu ở Buổi 1 và ít nhất một mô hình hồi quy). Các gói: mice, nanian, ggmmice, EValue.

Buổi	Nội dung	Sản phẩm
H1: Quy gán đa bội trong thực hành	EDA cho dữ liệu thiếu (các kiểu thiếu; so sánh người bệnh có và không có giá trị thiếu); mô hình quy gán; mice từng bước; chẩn đoán sự hội tụ và giá trị quy gán; gộp kết quả; ước lượng từ phân tích ca đầy đủ so với sau quy gán (FLUID-ICU Q4: lactate và tử vong 28 ngày)	Một phân tích quy gán kèm biểu đồ chẩn đoán và bảng so sánh
H2: Phân tích độ nhạy	Thay đổi gì và vì sao; các kịch bản MNAR (hiệu chỉnh delta); giá trị ngoại lai và điểm có ảnh hưởng lớn; các định nghĩa khác về phơi nhiễm và kết cục; nhiễu chưa đo lường và E-value; xác định trước các phân tích độ nhạy trong SAP; báo cáo chúng trong một bảng ngắn	Một bảng phân tích độ nhạy kèm hai câu diễn giải

5. Chọn học phần

Nếu bạn muốn...	Học phần
Tổng hợp toàn bộ bằng chứng về một câu hỏi	A: Phân tích gộp
Xây dựng hoặc đánh giá một thang điểm nguy cơ, công cụ dự báo hoặc mô hình AI	B: Dự báo và học máy
Kiểm tra một xét nghiệm, thang điểm hoặc dấu hiệu phát hiện bệnh tốt đến đâu	C: Độ chính xác chẩn đoán
Phân tích kết cục thời gian đến biến cố có biến cố cạnh tranh hoặc phơi nhiễm thay đổi	D: Phân tích sống còn nâng cao
Phân tích các số đo lặp lại theo thời gian trên cùng người bệnh	E: Dữ liệu dọc và mô hình hỗn hợp
Phân tích dữ liệu đa trung tâm, thử nghiệm ngẫu nhiên theo cụm hoặc sổ đăng ký	F: Dữ liệu phức tạp
Tạo một phân tích mà người khác chạy lại được, và biểu đồ đạt yêu cầu của tạp chí	G: Nghiên cứu có thể lặp lại (nên học đầu tiên)
Xử lý dữ liệu thiếu đúng cách và kiểm tra xem kết luận có vững hay không	H: Dữ liệu thiếu và phân tích độ nhạy

Các tài liệu miễn phí chính cho từng học phần được liệt kê trong `Course/References/further_reading_VN.md`: Cochrane Handbook và PRISMA 2020 (A); TRIPOD+AI và bài báo của Riley về cỡ mẫu (B); STARD 2015 và QUADAS-2 (C); *Statistics at Square One* và loạt bài Statistics Notes của Bland–Altman (D, E); CONSORT mở rộng cho thử nghiệm theo cụm (F); tài liệu hướng dẫn Quarto và hướng dẫn tác giả của tạp chí mục tiêu về biểu đồ (G); bài báo của Sterne trên *BMJ* năm 2009 về quy gán đa bội (H).