**TRƯỜNG ĐẠI HỌC Y TẾ CÔNG CỘNG  
 KHOA CÁC KHOA HỌC CƠ BẢN**



|  |  |
| --- | --- |
| **BÀI TẬP HẾT MÔN**  **Phân loại, mã hoá bệnh tật và tử vong**  **Lớp CNCQKHDL1-1A**  **NHÓM SỐ: 5**  **CHỦ ĐỀ: THIẾU MÁU DO THIẾU SẮT (ICD-10: D50)**   |  | | --- | | **Thành viên:**  **Nguyễn Hải An - 2211090001**  **Dương Mai Thu Hiền – 2211090012**  **Nguyễn Thị Thanh Nga - 2211090027**  **Đinh Lê Quỳnh Phương – 2211090031**    **Năm: 2025** |  MỤC LỤC [**MỤC LỤC 2**](#_Toc201320066)  [**DANH SÁCH BẢNG 7**](#_Toc201320067)  [**LỜI MỞ ĐẦU 7**](#_Toc201320068)  [**MỤC TIÊU VÀ PHƯƠNG PHÁP NGHIÊN CỨU 9**](#_Toc201320069)  [**1. Mục tiêu nghiên cứu: 9**](#_Toc201320070)  [**2. Phương pháp nghiên cứu 10**](#_Toc201320071)  [**I. Đọc và xử lý dữ liệu 11**](#_Toc201320072)  [**1.1. Nguồn dữ liệu 11**](#_Toc201320073)  [**1.2. Quy trình tiền xử lý 12**](#_Toc201320074)  [**1.2.1. Thực hiện bằng Python (Jupyter Notebook) 12**](#_Toc201320075)  [**1.2.2. Thực hiện bằng RMarkdown (ngôn ngữ R) 13**](#_Toc201320076)  [**II. Phân tích xu hướng và đặc điểm bệnh nhân thiếu máu ICD-10: D50 13**](#_Toc201320077)  [**2.1. Phân bố bệnh nhân theo nhóm tuổi 13**](#_Toc201320078)  [**2.2. Phân bố theo giới tính 15**](#_Toc201320079)  [**2.3. Phân bố loại thiếu máu thiếu sắt 16**](#_Toc201320080)  [**2.3.1. Phân bố theo loại thiếu máu thiếu sắt 16**](#_Toc201320081)  [**2.3.2. Phân bố giới tính theo loại thiếu máu 18**](#_Toc201320082)  [**2.3.3. Phân bố độ tuổi theo loại thiếu máu 21**](#_Toc201320083)  [**2.4. Phân tích thời gian nằm viện (LOS) 24**](#_Toc201320084)  [**2.4.1. Theo nhóm tuổi: 24**](#_Toc201320085)  [**2.4.1.1. Theo nhóm tuổi của Python 24**](#_Toc201320086)  [**2.4.1.3. Tổng hợp nhận xét chung (theo hướng tiểu luận) 26**](#_Toc201320087)  [**2.4.2. Theo giới tính: 28**](#_Toc201320088)  [**2.4.2.1. Theo giới tính của R 28**](#_Toc201320089)  [**2.4.2.2. Theo giới tính của Python 29**](#_Toc201320090)  [**2.4.2.3. Tổng hợp nhận xét chung 30**](#_Toc201320091)  [**2.4.3. Theo giới tính và theo nhóm tuổi: 31**](#_Toc201320092)  [**2.5. Phân tích đa biến 34**](#_Toc201320093)  [**2.5.1. Tương quan giữa tuổi và thời gian nằm viện 34**](#_Toc201320094)  [**2.5.1.1. Phân tán dữ liệu: 34**](#_Toc201320095)  [**2.5.1.2. Xu hướng tương quan: 34**](#_Toc201320096)  [**2.5.1.3. Sự hiện diện của các giá trị ngoại lệ và biến động lớn: 35**](#_Toc201320097)  [**2.5.1.4. Kết luận về tương quan: 35**](#_Toc201320098)  [**2.5.2. Phân tích ANOVA theo giới tính: 35**](#_Toc201320099)  [**2.6. Phân tích theo nguồn nhập viện 36**](#_Toc201320100)  [**2.6.1. Nguồn H (Recognised hospital - Bệnh viện được công nhận) là kênh nhập viện chủ yếu và quan trọng nhất: 37**](#_Toc201320101)  [**2.6.2. Nguồn T (Other - Nguồn khác) là kênh nhập viện thứ cấp đáng kể: 37**](#_Toc201320102)  [**2.6.3. Nguồn S (Other health care establishment - Cơ sở chăm sóc sức khỏe khác) và Nguồn N (Nursing home or other residential aged care - Viện dưỡng lão/cơ sở chăm sóc người cao tuổi nội trú khác) chiếm tỷ lệ rất nhỏ: 38**](#_Toc201320103)  [**2.6.4. Kết luận tổng thể về Nguồn nhập viện: 38**](#_Toc201320104)  [**2.7. Phân bố bệnh nhân theo nhóm tuổi 38**](#_Toc201320105)  [**2.7.1. Loại C (Newborn Admission - Trẻ sơ sinh nhập viện) là loại hình nhập viện chủ yếu: 40**](#_Toc201320106)  [**2.7.2. Loại X (Dữ liệu cũ) và Loại O (Posthumous Organ Procurement - Tiếp nhận để lấy tạng sau khi chết) chiếm tỷ lệ đáng kể: 40**](#_Toc201320107)  [**2.7.3. Các loại S (Hospital Transfer - Chuyển viện) và L (Statistical Admission - Nhập viện thống kê) chiếm tỷ lệ nhỏ: 41**](#_Toc201320108)  [**2.7.4. Thiếu loại M (Other Admission - Nhập viện thông thường): 41**](#_Toc201320109)  [**2.7.5. Kết luận tổng thể về Loại nhập viện: 41**](#_Toc201320110)  [**2.8. Phân tích theo DRG code ( mã nhóm bệnh) 42**](#_Toc201320111)  [**2.8.1. Mã I08A - Bệnh về tim mạch (Cardiovascular Diseases) là nhóm bệnh phổ biến nhất: 43**](#_Toc201320112)  [**2.8.2. Các mã phổ biến tiếp theo: 44**](#_Toc201320113)  [**2.8.3. Sự đa dạng của các mã DRG: 44**](#_Toc201320114)  [**2.8.4. Kết luận: 44**](#_Toc201320115)  [**2.9. Phân tích vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán 45**](#_Toc201320116)  [**2.9.1. Thiếu máu hiếm khi là chẩn đoán chính: 46**](#_Toc201320117)  [**2.9.2. Thiếu máu thường xuyên là chẩn đoán phụ, đặc biệt ở các vị trí 5, 6 và 7: 46**](#_Toc201320118)  [**2.9.3. Số lượng chẩn đoán thiếu máu giảm dần ở các vị trí phụ tiếp theo: 46**](#_Toc201320119)  [**2.9.4. Ý nghĩa lâm sàng và quản lý: 47**](#_Toc201320120)  [**2.9.5. Kết luận: 47**](#_Toc201320121)  [**III. PHÂN TÍCH NÂNG CAO 47**](#_Toc201320122)  [**3.1. Phân tích ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng 47**](#_Toc201320123)  [**3.1.1. Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R 47**](#_Toc201320124)  [**3.1.2. Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của Python 50**](#_Toc201320125)  [**3.2. Phân tích theo nhóm nguy cơ cao 54**](#_Toc201320126)  [**3.3. Xây dựng mô hình dự đoán thời gian nằm viện 55**](#_Toc201320127)  [**IV. Hệ thống báo cáo trực tuyến sử dụng Shiny 56**](#_Toc201320128)  [**4.1. Mục tiêu xây dựng 56**](#_Toc201320129)  [**4.2. Cấu trúc và công nghệ sử dụng 56**](#_Toc201320130)  [**4.3. Lý do xây dựng 2 phiên bản 57**](#_Toc201320131)  [**4.4. Lợi ích của hệ thống 57**](#_Toc201320132)  [**4.5. So sánh hai phiên bản hệ thống báo cáo Shiny 57**](#_Toc201320133)  [**4.6. Kết luận về lựa chọn triển khai 59**](#_Toc201320134)  [**V. Đánh giá chất lượng bộ dữ liệu 59**](#_Toc201320135)  [**VI. Lời khuyên dành cho người sử dụng dữ liệu thiếu máu do thiếu sắt 59**](#_Toc201320136)  [**THẢO LUẬN BÀI BÁO: DXPR – CÔNG CỤ CHUẨN HOÁ DỮ LIỆU ICD trong nghiên cứu HER 61**](#_Toc201320137)  [**I. Giới thiệu 61**](#_Toc201320138)  [**II. Chức năng chính của gói R dxpr 61**](#_Toc201320139)  [**2.1 Chuẩn hóa mã ICD 62**](#_Toc201320140)  [**2.2. Nhóm mã ICD 62**](#_Toc201320141)  [**III. Xử lý dữ liệu lâm sàng (Data Wrangling) 63**](#_Toc201320142)  [**IV. Xử lý dữ liệu thủ thuật (Procedure Data) 64**](#_Toc201320143)  [**V. Trực quan hóa và kiểm định 65**](#_Toc201320144)  [**VI. Phân tích hiệu suất (Performance) 65**](#_Toc201320145)  [**VII. Kết luận và hướng phát triển 66**](#_Toc201320146)  [**7.1. Ưu điểm: 66**](#_Toc201320147)  [**7.2. Hướng phát triển: 67**](#_Toc201320148)  [**VII. Tổng kết 68**](#_Toc201320149)  **DANH SÁCH HÌNH ẢNH**  [**Hình 1: Phân bố bệnh nhân theo nhóm tuổi thiếu máu do thiếu sắt 16**](#_Toc201317159)  [**Hình 2: Phân bố giới tính bệnh nhân thiếu máu di thiếu sắt 17**](#_Toc201317160)  [**Hình 3: Phân bố các loại thiếu máu do thiếu sắt 18**](#_Toc201317161)  [**Hình 4: Biểu đồ phân bố giới tính theo loại thiếu máu 20**](#_Toc201317162)  [**Hình 5: Biểu đồ phân bố độ tuổi theo loại máu 22**](#_Toc201317163)  [**Hình 6: Histogram phân phối thời gian nằm viện theo nhóm tuổi 25**](#_Toc201317164)  [**Hình 7: Boxplot thời gian nằm viện theo nhóm tuổi 25**](#_Toc201317165)  [**Hình 8: Histogram phân bố thời gian nằm viện trung bình theo nhóm tuổi 26**](#_Toc201317166)  [**Hình 9: Boxplot thời gian nằm viện trung bình theo nhóm tuổi 26**](#_Toc201317167)  [**Hình 10: Phân bố thời gian nằm viện theo giới tính" - Histogram song song 29**](#_Toc201317168)  [**Hình 11: "Phân bố thời gian nằm viện theo giới tính" - Boxplot với điểm trung bình 29**](#_Toc201317169)  [**Hình 12: "Phân phối thời gian nằm viện theo giới tính" (Kernel Density Estimate - KDE) 30**](#_Toc201317170)  [**Hình 13: Boxplot thời gian nằm viện theo giới tính 30**](#_Toc201317171)  [**Hình 14: Thời gian nằm viện trung bình theo nhóm tuổi và giới tính 33**](#_Toc201317172)  [**Hình 15: Tương quan giữa tuổi và thời gian nằm viện 35**](#_Toc201317173)  [**Hình 16: Biểu đồ tròn phân bố theo nguồn nhập viện 37**](#_Toc201317174)  [**Hình 17: Biểu đồ tròn phân bố theo loại nhập viện 40**](#_Toc201317175)  [**Hình 18: Top 10 mã nhóm bệnh phổ biến 43**](#_Toc201317176)  [**Hình 19: Top 20 mã nhóm bệnh phổ biến 44**](#_Toc201317177)  [**Hình 20: Phân bố vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán của Python 46**](#_Toc201317178)  [**Hình 21: Phân bố vị trí chẩn đoán thiếu máu trong danh sách của R 46**](#_Toc201317179)  [**Hình 22: Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R 49**](#_Toc201317180)  [**Hình 23: Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R 51**](#_Toc201317181)  [**Hình 24: Ma trận tương quan giữa LOS, tuổi và cân nặng 51**](#_Toc201317182)  [**Hình 25: Phân bố bệnh nhân theo nguy cơ 55**](#_Toc201317183)  [**Hình 26: Phân tích sai số dự đoán (residual) 56**](#_Toc201317184)  [**Hình 27: Dự đoán LOS vs Giá trị thực tế 56**](#_Toc201317185) DANH SÁCH BẢNG [**Bảng 1: Phân bố bệnh nhân theo nhóm tuổi thiếu máu do thiếu sắt 15**](#_Toc201317106)  [**Bảng 2: Phân bố các loại thiếu máu do thiếu sắt 18**](#_Toc201317107)  [**Bảng 3: Bảng phân bố giới tính theo loại thiếu máu 19**](#_Toc201317108)  [**Bảng 4: Bảng phân bố độ tuổi theo loại thiếu máu 22**](#_Toc201317109)  [**Bảng 5: Bảng phân bố theo nguồn nhập viện 37**](#_Toc201317110)  [**Bảng 6: Bảng phân bố theo loại nhập viện 40**](#_Toc201317111) |

# LỜI MỞ ĐẦU

Thiếu máu do thiếu sắt là một vấn đề sức khỏe cộng đồng toàn cầu, đặc biệt phổ biến tại các quốc gia đang phát triển. Đây là tình trạng xảy ra khi cơ thể không có đủ sắt để tổng hợp hemoglobin – protein thiết yếu trong hồng cầu giữ vai trò vận chuyển oxy đến các mô và cơ quan. Sự suy giảm số lượng và chất lượng hồng cầu không chỉ làm giảm chất lượng cuộc sống mà còn tác động sâu rộng đến nhiều hệ thống trong cơ thể như miễn dịch, thần kinh, đồng thời làm giảm khả năng hoạt động thể chất và nhận thức. Nguyên nhân chính thường do mất máu mạn tính, giảm hấp thu sắt hoặc chế độ dinh dưỡng thiếu hụt. Nếu không được chẩn đoán và điều trị kịp thời, tình trạng này có thể dẫn đến nhiều biến chứng nghiêm trọng, ảnh hưởng lâu dài đến sức khỏe người bệnh.

Trong bối cảnh đó, việc phân tích dữ liệu chuyên sâu về các nhóm bệnh nhân thiếu máu do thiếu sắt đóng vai trò then chốt trong việc hiểu rõ đặc điểm dịch tễ, từ đó hỗ trợ cải thiện chất lượng chăm sóc y tế và phân bổ nguồn lực hiệu quả. Theo Phân loại thống kê quốc tế về bệnh tật (ICD-10), thiếu máu do thiếu sắt được mã hóa thành các nhóm chính: D50.0 (thiếu máu do mất máu mạn tính), D50.8 (các loại thiếu máu do thiếu sắt khác), và D50.9 (thiếu máu do thiếu sắt không xác định).

Trước thực trạng và tầm quan trọng của vấn đề, nhóm chúng tôi thực hiện đề tài nghiên cứu tập trung vào việc phân tích dữ liệu bệnh nhân được chẩn đoán thiếu máu do thiếu sắt (nhóm mã ICD-10: D50). Mục tiêu của nghiên cứu là cung cấp một cái nhìn toàn diện về đặc điểm của nhóm bệnh nhân này, làm cơ sở khoa học cho việc xây dựng các giải pháp can thiệp hiệu quả. Các mục tiêu cụ thể bao gồm:

1. **Mô tả đặc điểm nhân khẩu học** (tuổi, giới tính) và phân bố bệnh nhân theo từng mã chẩn đoán cụ thể (D50.0, D50.8, D50.9).

2. **Đánh giá thời gian nằm viện (LOS)** và phân tích các yếu tố liên quan như tuổi, giới tính và nguồn nhập viện.

3. **Phân tích các bệnh lý đi kèm** và vị trí của chẩn đoán trong hồ sơ bệnh án để hiểu rõ hơn về gánh nặng bệnh tật.

4. **Xây dựng mô hình dự đoán** thời gian nằm viện dựa trên các biến số lâm sàng sẵn có.

Để thực hiện các mục tiêu trên, nghiên cứu sử dụng bộ dữ liệu được trích xuất từ hồ sơ bệnh án điện tử, bao gồm Admission.xlsx (thông tin nhập viện) và Diagnosis.xlsx (thông tin chẩn đoán). Dữ liệu sẽ được tiền xử lý và phân tích bằng ngôn ngữ Python với thư viện Pandas, sau đó kết quả sẽ được trực quan hóa bằng R Markdown và Shiny. Kết quả phân tích không chỉ mang ý nghĩa thống kê mà còn có giá trị thực tiễn, góp phần tối ưu hóa quy trình điều trị, nâng cao chất lượng chăm sóc bệnh nhân thiếu máu do thiếu sắt và tăng cường hiểu biết về tình trạng sức khỏe cộng đồng quan trọng này.

# MỤC TIÊU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

# Mục tiêu nghiên cứu:

* Phân tích thực trạng mắc bệnh thiếu máu do thiếu sắt (ICD-10: D50.0, D50.8, D50.9) trong cộng đồng/hệ thống bệnh viện thông qua bộ dữ liệu nhập viện và chẩn đoán.
* Xác định đặc điểm nhân khẩu học (tuổi, giới tính), mô hình phân bố bệnh theo thời gian và theo nhóm tuổi.
* Phân tích các bệnh kèm theo (comorbidities) thường gặp ở bệnh nhân thiếu máu do thiếu sắt.
* Hỗ trợ đề xuất các biện pháp can thiệp hoặc xây dựng kế hoạch phòng ngừa/chăm sóc sức khỏe phù hợp với các nhóm nguy cơ.

# Phương pháp nghiên cứu

* **Phương pháp nghiên cứu**: Mô tả hồi cứu (retrospective descriptive analysis) dựa trên bộ dữ liệu bệnh viện được cung cấp.
* **Dữ liệu sử dụng**:
  1. **admission**: Thông tin nhập viện (mã bệnh nhân, tuổi, giới, nguồn nhập viện, số ngày nằm viện…).
  2. **diagnosis**: Chẩn đoán bệnh của từng lượt nhập viện.
  3. **ICD10**: Danh mục mã hóa bệnh theo tiêu chuẩn ICD-10.
* **Quy trình phân tích**:
  1. Lọc các bệnh nhân có mã chẩn đoán D50.0, D50.8, D50.9.
  2. Gộp thông tin từ các bảng để có cái nhìn tổng quát về mỗi lượt nhập viện liên quan đến thiếu máu thiếu sắt.
  3. Phân nhóm theo độ tuổi và giới tính để mô tả đặc điểm bệnh nhân.
  4. Phân tích thời gian nhập viện (năm, tháng) để phát hiện xu hướng theo thời gian.
  5. Xác định các bệnh kèm theo thường gặp trong các trường hợp thiếu máu.
* **Công cụ phân tích**: Ngôn ngữ lập trình **R** và **Python**, sử dụng các thư viện như pandas, ggplot2, dplyr, seaborn, matplotlib.

# 

# I. Đọc và xử lý dữ liệu

# 1.1. Nguồn dữ liệu

Dữ liệu nghiên cứu được trích xuất từ **hồ sơ bệnh án điện tử**, bao gồm ba bảng dữ liệu chính:

* **admission.xlsx**: Cung cấp thông tin hành chính và lâm sàng liên quan đến quá trình nhập viện, bao gồm mã bệnh nhân, độ tuổi, giới tính, thời gian nằm viện và nguồn vào viện.
* **diagnosis.xlsx**: Chứa danh sách các mã chẩn đoán bệnh theo hệ thống **ICD-10**, dùng để xác định các trường hợp thiếu máu do thiếu sắt.
* **ICD10.xlsx**: Danh mục mã ICD-10 kèm theo mô tả chi tiết, giúp giải nghĩa và phân nhóm các bệnh lý trong phân tích.

# 1.2. Quy trình tiền xử lý

Nhằm phục vụ mục tiêu phân tích bệnh thiếu máu do thiếu sắt (ICD-10: D50.0, D50.8, D50.9), nhóm đã thực hiện các bước xử lý dữ liệu như sau:

## 1.2.1. Thực hiện bằng Python (Jupyter Notebook)

* **Bước 1**: Đọc dữ liệu từ ba bảng: admission, diagnosis, và ICD10 từ các tệp Excel.
* **Bước 2**: Lọc ra các dòng trong bảng diagnosis có mã ICD-10 thuộc nhóm **thiếu máu do thiếu sắt**, bao gồm:
* D50.0: Thiếu máu do mất máu mạn tính
* D50.8: Thiếu máu do nguyên nhân khác
* D50.9: Thiếu máu thiếu sắt không xác định nguyên nhân
* **Bước 3**: Trích danh sách admission\_id tương ứng từ các dòng chẩn đoán trên để làm khóa lọc dữ liệu liên quan trong bảng admission.
* **Bước 4**: Lọc lại hai bảng admission và diagnosis, chỉ giữ lại những dòng có admission\_id thuộc nhóm bệnh nhân đã xác định.
* **Bước 5**: Kiểm tra tính nhất quán giữa các bảng dữ liệu sau khi lọc, đảm bảo không có thiếu sót, trùng lặp hoặc sai lệch.
* **Bước 6**: Lưu kết quả lọc vào hai tệp:
* filtered\_admission.xlsx: Danh sách các lượt nhập viện của bệnh nhân thiếu máu do thiếu sắt.
* diagnosis\_chi\_D500\_D508\_D509.xlsx: Thông tin chẩn đoán phù hợp với nhóm bệnh mục tiêu.

## 1.2.2. Thực hiện bằng RMarkdown (ngôn ngữ R)

Quy trình xử lý dữ liệu sẽ tương tự, bao gồm:

* **Bước 1**: Đọc ba bảng dữ liệu từ các file .xlsx bằng các hàm như read\_excel() trong thư viện readxl.
* **Bước 2**: Lọc các dòng trong bảng diagnosis có mã ICD-10 nằm trong nhóm D50.0, D50.8, D50.9.
* **Bước 3**: Trích admission\_id từ các chẩn đoán phù hợp để làm khóa lọc dữ liệu từ bảng admission.
* **Bước 4**: Dùng hàm filter() trong dplyr để giữ lại các dòng có liên quan đến nhóm bệnh nhân mục tiêu.
* **Bước 5**: Kiểm tra dữ liệu sau lọc bằng các hàm như nrow(), distinct(), hoặc anti\_join() để xác minh tính đầy đủ và đồng nhất.
* **Bước 6**: Xuất dữ liệu đã xử lý ra tệp .xlsx bằng các gói như writexl hoặc openxlsx.

# II. Phân tích xu hướng và đặc điểm bệnh nhân thiếu máu ICD-10: D50

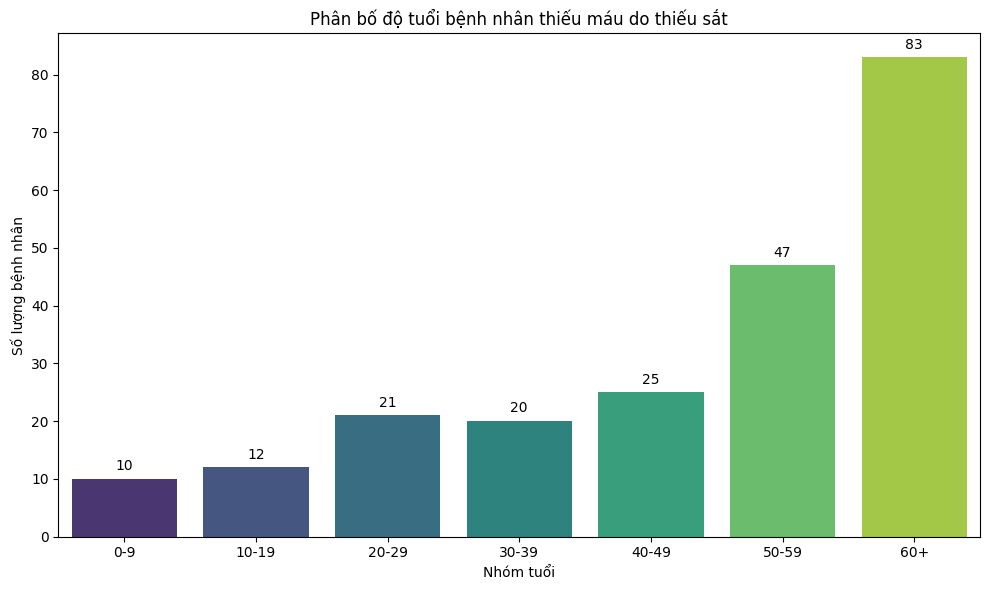
Sau bước tiền xử lý, nhóm tiến hành phân tích mô tả trên tập dữ liệu merged\_df – được tạo bằng cách gộp 3 bảng diagnosis, admission, và ICD10. Phân tích mô tả nhằm cung cấp cái nhìn tổng quan về đặc điểm nhân khẩu học và phân bố của bệnh thiếu máu do thiếu sắt trong dân số nghiên cứu.

# 2.1. Phân bố bệnh nhân theo nhóm tuổi

Dữ liệu tuổi (age\_years) của bệnh nhân được phân nhóm thành 7 nhóm: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, và ≥60 tuổi. Việc phân nhóm này giúp thuận tiện trong việc thống kê, so sánh và trình bày số liệu.

Bảng 1: Phân bố bệnh nhân theo nhóm tuổi thiếu máu do thiếu sắt

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Nhóm tuổi** | **Số lượng bệnh nhân** | **Tỷ lệ** |
| 0 | 0-9 | 10 | 4.59 |
| 1 | 10-19 | 12 | 5.50 |
| 2 | 20-29 | 21 | 9.63 |
| 3 | 30-39 | 20 | 9.17 |
| 4 | 40-49 | 25 | 11.47 |
| 5 | 50-59 | 47 | 21.56 |
| 6 | 60+ | 83 | 38.07 |



Hình 1: Phân bố bệnh nhân theo nhóm tuổi thiếu máu do thiếu sắt

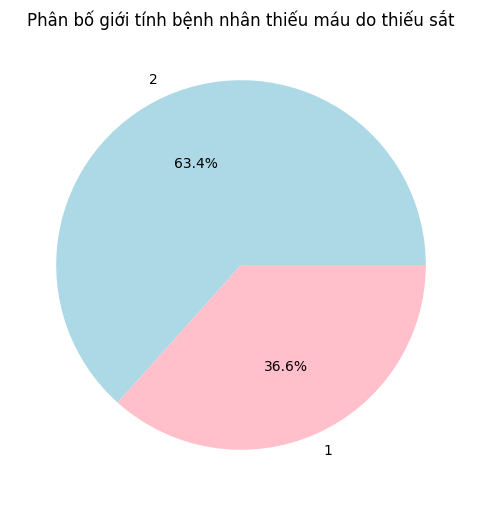
Nhận xét: Phân tích phân bố bệnh nhân theo nhóm tuổi cho thấy một xu hướng đáng chú ý: **nhóm từ 60 tuổi trở lên chiếm tỷ lệ cao nhất với 83 bệnh nhân**, tiếp theo là **nhóm 50–59 tuổi với 47 bệnh nhân**. Đây là các nhóm tuổi lớn, thường đi kèm với nhiều yếu tố nguy cơ như suy giảm khả năng hấp thu chất dinh dưỡng, tình trạng dinh dưỡng kém, và các bệnh lý mạn tính – những nguyên nhân phổ biến dẫn đến thiếu máu thiếu sắt.

Ngoài ra, **nhóm tuổi 20–29 cũng ghi nhận số lượng ca mắc tương đối cao**. Điều này hoàn toàn phù hợp với đặc điểm dịch tễ học của thiếu máu thiếu sắt ở **phụ nữ trong độ tuổi sinh sản**, do nhu cầu sắt tăng cao và nguy cơ mất máu qua kinh nguyệt.

# 2.2. Phân bố theo giới tính

Dữ liệu sau xử lý cho thấy phân bố giới tính bệnh nhân mắc thiếu máu do thiếu sắt như sau:

* Nữ giới chiếm 63.4%
* Nam giới chiếm 36.6%



Hình 2: Phân bố giới tính bệnh nhân thiếu máu di thiếu sắt

Nhận xét: Tỷ lệ này cho thấy **phụ nữ có nguy cơ thiếu máu cao hơn nam giới**, điều này phù hợp với thực tế lâm sàng. Nguyên nhân có thể đến từ **kinh nguyệt kéo dài, mang thai, cho con bú hoặc chế độ ăn thiếu sắt**.

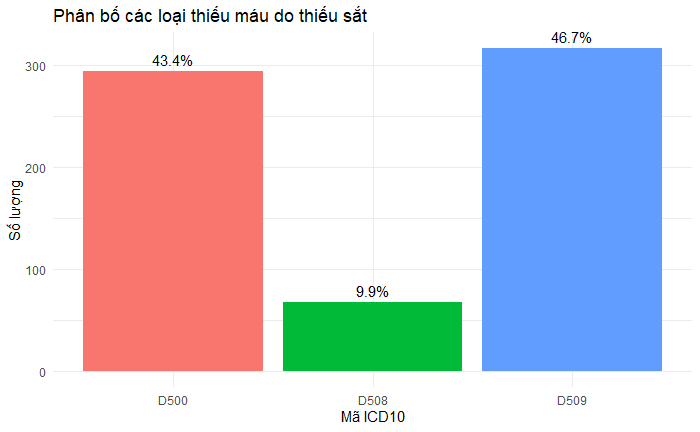
Do đó, phụ nữ – đặc biệt là trong độ tuổi sinh sản – cần được ưu tiên trong công tác sàng lọc, tư vấn và dự phòng thiếu máu.

# Phân bố loại thiếu máu thiếu sắt

## Phân bố theo loại thiếu máu thiếu sắt

Bảng 2: Phân bố các loại thiếu máu do thiếu sắt

| **Mã ICD-10** | **Số lượng bệnh nhân** | **Tỷ lệ (%)** |
| --- | --- | --- |
| D50.9 | 316 | 46.7 |
| D50.8 | 67 | 9.9 |
| D50.0 | 294 | 43.4 |



Hình 3: Phân bố các loại thiếu máu do thiếu sắt

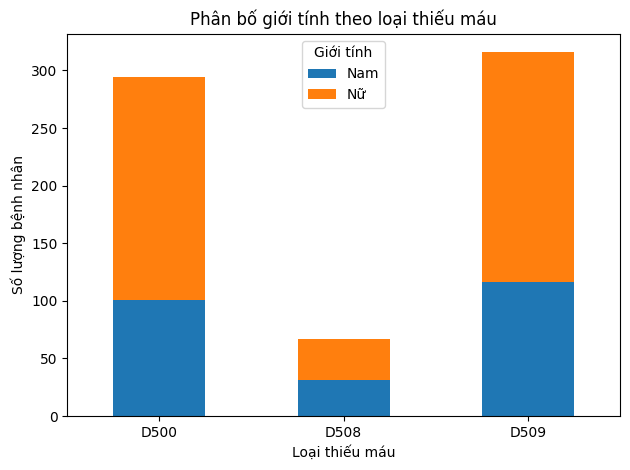
Nhận xét:

* Thiếu máu không xác định nguyên nhân (D50.9) chiếm tỷ lệ cao nhất với 46.7%, phản ánh thực tế lâm sàng là có nhiều trường hợp không xác định được nguyên nhân cụ thể gây thiếu máu do thiếu sắt. Điều này thường gặp ở bệnh nhân lớn tuổi, có đa bệnh lý hoặc không được thực hiện đầy đủ các xét nghiệm chuyên sâu.
* Thiếu máu do mất máu mạn tính (D50.0) đứng thứ hai với 43.4%, cho thấy đây là một nguyên nhân quan trọng, thường gặp ở các bệnh nhân có xuất huyết tiêu hóa (như loét dạ dày tá tràng, trĩ, polyp đại trực tràng) hoặc xuất huyết phụ khoa (như rong kinh, u xơ tử cung) – đặc biệt phổ biến ở phụ nữ trung niên.
* Thiếu máu do nguyên nhân khác (D50.8) chỉ chiếm 9.9%, cho thấy các nguyên nhân ít phổ biến hơn như rối loạn hấp thu sắt (bệnh Celiac, viêm ruột mạn), bệnh mạn tính gây ảnh hưởng chuyển hóa sắt không phải là cơ chế chính trong nhóm bệnh nhân được nghiên cứu.

## Phân bố giới tính theo loại thiếu máu

Bảng 3: Bảng phân bố giới tính theo loại thiếu máu

| **Gender** | **Nam** | **Nữ** |
| --- | --- | --- |
| D500 | 101 | 193 |
| D508 | 31 | 36 |
| D509 | 116 | 200 |

Nhận xét:

Hình 4: Biểu đồ phân bố giới tính theo loại thiếu máu

Biểu đồ cột chồng "Phân bố giới tính theo loại thiếu máu" cùng với bảng số liệu chi tiết minh họa số lượng bệnh nhân Nam và Nữ mắc ba loại thiếu máu cụ thể thuộc nhóm D50: thiếu máu do thiếu sắt, với định nghĩa chính xác từ ICD-10.

**1. Nữ giới chiếm ưu thế rõ rệt trong tất cả các loại thiếu máu do thiếu sắt:**

* **Đối với D50.0 (Thiếu máu do thiếu sắt thứ phát sau mất máu mạn tính):** Có 101 bệnh nhân Nam và 193 bệnh nhân Nữ.
* **Đối với D50.8 (Thiếu máu do thiếu sắt khác):** Có 31 bệnh nhân Nam và 36 bệnh nhân Nữ.
* **Đối với D50.9 (Thiếu máu do thiếu sắt không xác định):** Có 116 bệnh nhân Nam và 200 bệnh nhân Nữ.
* Trong cả ba phân loại thiếu máu do thiếu sắt này, số lượng bệnh nhân **Nữ giới luôn cao hơn đáng kể** so với Nam giới. Điều này đặc biệt nổi bật ở D50.0 và D50.9, nơi số lượng Nữ giới gần gấp đôi hoặc hơn Nam giới. Đây là một phát hiện phù hợp với các yếu tố sinh lý học và bệnh lý liên quan đến giới tính, đặc biệt là nguy cơ mất máu mạn tính (ví dụ do kinh nguyệt, sinh nở) ở phụ nữ, dẫn đến thiếu sắt.

**2. D50.0 và D50.9 là các loại thiếu máu do thiếu sắt phổ biến nhất:**

* Tổng số bệnh nhân mắc **D50.0 (Thiếu máu do thiếu sắt thứ phát sau mất máu mạn tính)** là 294 người (101 Nam + 193 Nữ).
* Tổng số bệnh nhân mắc **D50.9 (Thiếu máu do thiếu sắt không xác định)** là 316 người (116 Nam + 200 Nữ).
* **D50.8 (Thiếu máu do thiếu sắt khác)** có số lượng thấp nhất với tổng cộng 67 người (31 Nam + 36 Nữ).
* Sự phổ biến của D50.0 cho thấy mất máu mạn tính là một nguyên nhân đáng kể gây thiếu máu do thiếu sắt trong tập dữ liệu này. Tỷ lệ cao của D50.9 có thể phản ánh việc trong nhiều trường hợp, nguyên nhân cụ thể của thiếu máu do thiếu sắt chưa được xác định rõ ràng hoặc không được ghi nhận chi tiết, hoặc nó là một chẩn đoán ban đầu trước khi có thể xác định nguyên nhân sâu hơn.

**3. Sự chênh lệch giới tính nổi bật nhất ở D50.0 và D50.9:**

* Khoảng cách về số lượng bệnh nhân giữa Nữ và Nam là lớn nhất ở D50.0 và D50.9. Điều này một lần nữa củng cố nhận định về gánh nặng thiếu máu do thiếu sắt lớn hơn ở phụ nữ, đặc biệt là các trường hợp liên quan đến mất máu mạn tính hoặc các trường hợp chưa xác định được nguyên nhân cụ thể.

**Kết luận và Hàm ý Y tế:**

Phân tích này, được hỗ trợ bởi định nghĩa chính xác từ ICD-10, làm nổi bật rằng **thiếu máu do thiếu sắt là một vấn đề sức khỏe ưu thế ở Nữ giới** trong tập dữ liệu này. Đặc biệt, các trường hợp **thiếu máu do thiếu sắt thứ phát sau mất máu mạn tính (D50.0)** và các trường hợp **không xác định nguyên nhân cụ thể (D50.9)** là những loại phổ biến nhất.

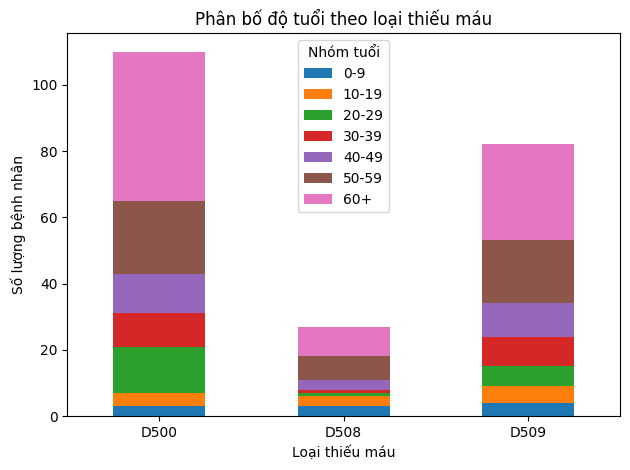
**Hàm ý lâm sàng và quản lý sức khỏe cộng đồng:**

* **Phụ nữ là nhóm đối tượng nguy cơ cao:** Các chương trình sàng lọc, dự phòng và điều trị thiếu máu do thiếu sắt cần đặc biệt chú trọng đến phụ nữ, bao gồm giáo dục về chế độ dinh dưỡng, bổ sung sắt, và quản lý các tình trạng gây mất máu mạn tính.
* **Nhu cầu chẩn đoán sâu hơn:** Tỷ lệ đáng kể của D50.9 cho thấy có thể cần khuyến khích các nỗ lực chẩn đoán sâu hơn để xác định nguyên nhân gốc rễ của thiếu máu do thiếu sắt, đặc biệt khi nguyên nhân ban đầu chưa rõ ràng, nhằm đưa ra phác đồ điều trị tối ưu và phòng ngừa tái phát.
* **Quản lý mất máu mạn tính:** Các bệnh viện và cơ sở y tế cần quan tâm đặc biệt đến việc phát hiện và quản lý hiệu quả các tình trạng gây mất máu mạn tính để giảm thiểu tỷ lệ thiếu máu do thiếu sắt.

## Phân bố độ tuổi theo loại thiếu máu

Bảng 4: Bảng phân bố độ tuổi theo loại thiếu máu

| **age\_group** | **0-9** | **10-19** | **20-29** | **30-39** | **40-49** | **50-59** | **60+** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **D500** | 3 | 4 | 14 | 10 | 12 | 22 | 45 |
| **D508** | 3 | 3 | 1 | 1 | 3 | 7 | 9 |
| **D509** | 4 | 5 | 6 | 9 | 10 | 19 | 29 |



Hình 5: Biểu đồ phân bố độ tuổi theo loại máu

**Nhận xét về Phân bố các loại thiếu máu do thiếu sắt (D500, D508, D509) theo nhóm tuổi**

Bảng số liệu này cung cấp cái nhìn chi tiết về số lượng bệnh nhân mắc ba loại thiếu máu do thiếu sắt (D50.0: Thiếu máu do thiếu sắt thứ phát sau mất máu mạn tính; D50.8: Thiếu máu do thiếu sắt khác; D50.9: Thiếu máu do thiếu sắt không xác định) trong từng nhóm tuổi.

**1. Thiếu máu do thiếu sắt (D50.0, D50.8, D50.9) có xu hướng tăng theo tuổi:**

* Đối với cả ba loại thiếu máu, số lượng bệnh nhân có xu hướng tăng dần theo các nhóm tuổi cao hơn.
* **Đặc biệt rõ rệt ở nhóm tuổi 60+:** Nhóm này có số lượng bệnh nhân mắc D50.0 (45 trường hợp), D50.8 (9 trường hợp), và D50.9 (29 trường hợp) cao nhất so với tất cả các nhóm tuổi khác.
* **Nhóm 50-59 tuổi** cũng có số lượng đáng kể, đứng thứ hai sau nhóm 60+.
* Điều này phù hợp với thực tế lâm sàng, khi người lớn tuổi thường có nhiều bệnh lý nền, chế độ dinh dưỡng kém, hoặc các tình trạng mất máu mạn tính (ví dụ: xuất huyết tiêu hóa, bệnh thận mạn tính) làm tăng nguy cơ thiếu máu do thiếu sắt.

**2. Các nhóm tuổi trẻ có tỷ lệ mắc thiếu máu do thiếu sắt thấp hơn:**

* Các nhóm tuổi từ **0-9, 10-19, 20-29** đều có số lượng bệnh nhân mắc cả ba loại thiếu máu thấp hơn nhiều so với các nhóm tuổi trung niên và cao tuổi.
* Tuy nhiên, vẫn có một số trường hợp thiếu máu do thiếu sắt được ghi nhận ở trẻ em (nhóm 0-9), đặc biệt là D50.0 (3 trường hợp) và D50.9 (4 trường hợp), điều này có thể liên quan đến yếu tố dinh dưỡng hoặc các bệnh lý bẩm sinh.

**3. Phân bố cụ thể theo từng loại thiếu máu:**

* **D50.0 (Thiếu máu do thiếu sắt thứ phát sau mất máu mạn tính):**
  + Số ca mắc tăng dần và đạt đỉnh ở nhóm 60+ (45 ca).
  + Điều này gợi ý rằng mất máu mạn tính là một nguyên nhân quan trọng gây thiếu máu do thiếu sắt ở người lớn tuổi.
* **D50.8 (Thiếu máu do thiếu sắt khác):**
  + Có số lượng thấp nhất trong ba loại và cũng tăng theo tuổi, với đỉnh ở nhóm 60+ (9 ca).
  + Loại này có thể bao gồm các nguyên nhân ít phổ biến hơn hoặc chưa được phân loại cụ thể vào D50.0 hay D50.9.
* **D50.9 (Thiếu máu do thiếu sắt không xác định):**
  + Cũng cho thấy xu hướng tăng theo tuổi, với nhóm 60+ có số lượng cao nhất (29 ca).
  + Số lượng cao ở loại "không xác định" này, đặc biệt ở người lớn tuổi, có thể gợi ý rằng việc chẩn đoán nguyên nhân cụ thể của thiếu máu do thiếu sắt ở nhóm này có thể gặp khó khăn hoặc chưa được thực hiện đầy đủ.

**Kết luận và Hàm ý:**

Bảng phân bố này khẳng định rằng **thiếu máu do thiếu sắt, đặc biệt là các loại liên quan đến mất máu mạn tính (D50.0) và các trường hợp chưa xác định nguyên nhân (D50.9), là một vấn đề sức khỏe có tỷ lệ mắc cao hơn đáng kể ở các nhóm tuổi lớn hơn**, đặc biệt là từ 50 tuổi trở lên và cao nhất ở nhóm 60+.

**Hàm ý y tế:**

* **Sàng lọc và quản lý ở người cao tuổi:** Các chương trình sàng lọc thiếu máu do thiếu sắt cần được chú trọng ở người lớn tuổi, và khi phát hiện, cần tìm kiếm kỹ lưỡng các nguyên nhân tiềm ẩn, đặc biệt là mất máu mạn tính.
* **Tối ưu hóa chẩn đoán:** Cần tăng cường nỗ lực để xác định nguyên nhân cụ thể của thiếu máu do thiếu sắt, giảm số lượng chẩn đoán D50.9, từ đó giúp đưa ra phác đồ điều trị hiệu quả và chính xác hơn cho bệnh nhân.
* **Giáo dục sức khỏe:** Nâng cao nhận thức về các yếu tố nguy cơ và dấu hiệu thiếu máu do thiếu sắt ở người lớn tuổi cũng là một việc quan trọng.

# 2.4. Phân tích thời gian nằm viện (LOS)

## 2.4.1. Theo nhóm tuổi:

### **2.4.1.1. Theo nhóm tuổi của Python**

Hình 6: Histogram phân phối thời gian nằm viện theo nhóm tuổi

Hình 7: Boxplot thời gian nằm viện theo nhóm tuổi

**Nhận xét chung:** Hai biểu đồ này rất hiệu quả trong việc minh họa phân phối và sự hiện diện của các giá trị ngoại lệ theo từng nhóm tuổi.

**Hình 6 ("Phân phối thời gian nằm viện theo nhóm tuổi" - Histogram):**

**- Điểm nổi bật:** Biểu đồ này cho thấy rõ ràng phân bố số lượng bệnh nhân theo thời gian nằm viện và theo nhóm tuổi. Phần lớn bệnh nhân ở mọi nhóm tuổi đều có thời gian nằm viện ngắn (dưới 10-20 ngày).

**- Các nhóm có LOS dài:** Nhóm 60+ (màu hồng/tím) có số lượng bệnh nhân tập trung ở thời gian nằm viện ngắn nhiều nhất, nhưng cũng có một số bệnh nhân kéo dài đến hơn 50 ngày, thậm chí có những trường hợp rất hiếm lên tới trên 500 ngày. Điều này gợi ý rằng nhóm 60+ có thể có xu hướng nằm viện lâu hơn trong một số trường hợp. Nhóm 30-39 (màu đỏ) cũng có một số trường hợp kéo dài, nhưng khó đánh giá cụ thể mức độ ảnh hưởng của chúng đến trung bình chỉ qua biểu đồ này.

**- Tính trực quan:** Biểu đồ chồng rất tốt để so sánh số lượng bệnh nhân và phân bố trong các khoảng thời gian nằm viện theo từng nhóm tuổi.

**Hình 7 ("Boxplot thời gian nằm viện theo nhóm tuổi"):**

**- Điểm nổi bật:** Biểu đồ này cực kỳ hữu ích để xác định giá trị trung vị (median), khoảng tứ phân vị (IQR) và đặc biệt là các **giá trị ngoại lệ (outliers)** cho từng nhóm tuổi.

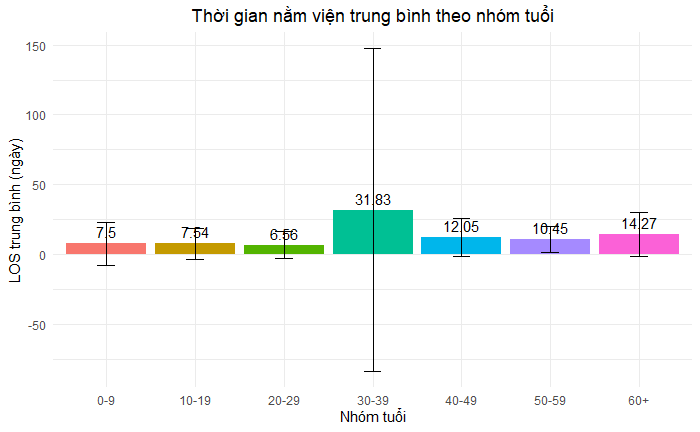
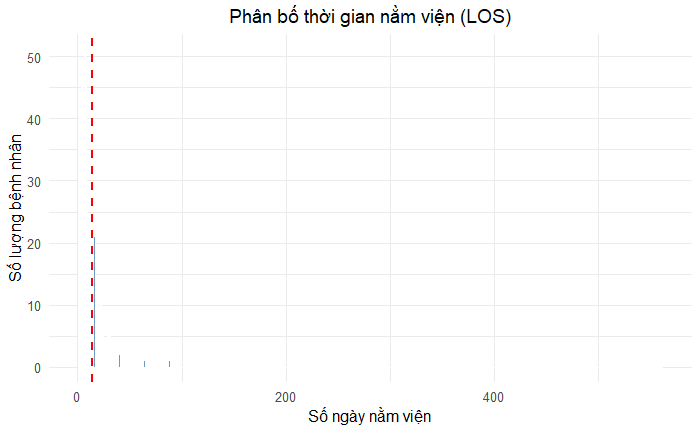
**- Outliers đáng chú ý:**

+ Nhóm **30-39** có một giá trị ngoại lệ rất lớn, vượt quá 550 ngày. Giá trị này cực đoan đến mức nó một mình có thể kéo giá trị trung bình của nhóm lên rất cao, phù hợp với số liệu 31.83 ngày trung bình từ biểu đồ R.

+ Nhóm **60+** cũng có nhiều giá trị ngoại lệ kéo dài đến khoảng 100 ngày và một vài trường hợp vượt quá 100 ngày. Mặc dù không có giá trị nào cao bằng outliers của nhóm 30-39, số lượng outliers nhiều hơn ở nhóm 60+ cũng có thể ảnh hưởng lớn đến trung bình của nhóm này.

**- Median và IQR:** Các nhóm tuổi trẻ có median và IQR rất thấp, cho thấy phần lớn bệnh nhân nằm viện ngắn. Ngược lại, nhóm 60+ có median và IQR cao hơn một chút so với các nhóm trẻ, cho thấy một xu hướng chung về thời gian nằm viện dài hơn.

**- Tính trực quan:** Biểu đồ hộp là công cụ tuyệt vời để xem xét phân bố, sự lệch và sự hiện diện của các giá trị bất thường một cách nhanh chóng.

**2.4.1.2. Theo nhóm tuổi của R**

Hình 9: Boxplot thời gian nằm viện trung bình theo nhóm tuổi

Hình 8: Histogram phân bố thời gian nằm viện trung bình theo nhóm tuổi

**Nhận xét chung:** Hai biểu đồ này cung cấp cái nhìn chi tiết về thời gian nằm viện, đặc biệt là giá trị trung bình và phân bố tổng thể.

**Hình 8 ("Histogram phân bố thời gian nằm viện trung bình theo nhóm tuổi"):**

**- Điểm nổi bật:** Nhóm tuổi **30-39** có thời gian nằm viện trung bình cao nhất một cách đáng ngạc nhiên (31.83 ngày). Điều này đi kèm với một thanh lỗi (error bar) rất lớn, cho thấy sự biến động cực kỳ cao trong thời gian nằm viện của nhóm này.

**- Các nhóm khác:** Nhóm 60+ có trung bình 14.27 ngày, cao hơn các nhóm trẻ nhưng thấp hơn nhóm 30-39. Các nhóm 40-49 và 50-59 có trung bình lần lượt là 12.05 và 10.45 ngày. Các nhóm trẻ nhất (0-9, 10-19, 20-29) có thời gian nằm viện trung bình thấp và tương đối ổn định (quanh 6-7 ngày).

**- Tính trực quan:** Biểu đồ này trực tiếp hiển thị giá trị trung bình, rất dễ đọc và so sánh giữa các nhóm. Tuy nhiên, cột lỗi cực lớn ở nhóm 30-39 là một điểm cần được chú ý và điều tra sâu hơn về nguyên nhân.

**Hình 9 ("Boxplot thời gian nằm viện trung bình theo nhóm tuổi”)**

**- Điểm nổi bật:** Biểu đồ này cho thấy phần lớn bệnh nhân có thời gian nằm viện rất ngắn, tập trung sát trục Y (0-khoảng 10 ngày). Đường đứt khúc màu đỏ tại gần điểm 0 nhấn mạnh điều này.

**- Tính phân tán:** Có một số lượng nhỏ bệnh nhân có thời gian nằm viện rất dài, trải dài ra xa trục Y, thậm chí vượt quá 400 ngày. Điều này xác nhận sự tồn tại của các trường hợp ngoại lệ với thời gian nằm viện cực đoan, làm ảnh hưởng đến giá trị trung bình tổng thể.

**- Tính trực quan:** Biểu đồ này trực quan về việc nhìn thấy sự tập trung của dữ liệu ở khoảng thời gian ngắn và sự phân tán của các giá trị lớn. Tuy nhiên, do không có phân loại theo nhóm tuổi, nó chỉ cung cấp cái nhìn tổng quát mà không chỉ ra nhóm tuổi cụ thể nào chịu ảnh hưởng bởi các trường hợp nằm viện dài này.

### **2.4.1.3. Tổng hợp nhận xét chung (theo hướng tiểu luận)**

Dựa trên phân tích toàn diện từ bốn biểu đồ về thời gian nằm viện (Length of Stay - LOS) theo nhóm tuổi, chúng ta có thể rút ra những nhận định quan trọng sau:

**1. Phân bố tổng thể thời gian nằm viện:** Các biểu đồ đều chỉ ra rằng phần lớn bệnh nhân có thời gian nằm viện tương đối ngắn, thường dưới 10-20 ngày (thể hiện rõ trên biểu đồ phân bố tổng thể và biểu đồ chồng). Tuy nhiên, có một tỷ lệ nhỏ các trường hợp đặc biệt với thời gian nằm viện kéo dài bất thường, đôi khi lên tới vài trăm ngày.

**2. Nhóm tuổi có thời gian nằm viện trung bình cao nhất:** Trái ngược với những suy luận ban đầu về việc nhóm tuổi cao nhất sẽ có LOS trung bình dài nhất, **nhóm tuổi 30-39 nổi bật là nhóm có thời gian nằm viện trung bình cao nhất, đạt 31.83 ngày.** Phân tích từ biểu đồ hộp cho thấy điều này chủ yếu là do sự hiện diện của một hoặc một vài trường hợp ngoại lệ cực đoan (outlier) với thời gian nằm viện rất dài (có thể lên tới hơn 550 ngày). Cột lỗi rất lớn đi kèm trong biểu đồ trung bình cũng minh chứng cho sự biến động mạnh mẽ này.

**3. Thời gian nằm viện trung bình của nhóm cao tuổi (60+):** Nhóm tuổi 60+ có thời gian nằm viện trung bình là 14.27 ngày. Mặc dù không phải là cao nhất, nhưng con số này cao hơn đáng kể so với các nhóm tuổi trẻ hơn (ví dụ, gấp khoảng 2 lần so với nhóm 20-29 có trung bình 6.56 ngày). Biểu đồ hộp và biểu đồ chồng cũng cho thấy nhóm 60+ có nhiều trường hợp ngoại lệ và phân bố thời gian nằm viện kéo dài hơn so với các nhóm trẻ hơn, mặc dù không có trường hợp nào cực đoan như ở nhóm 30-39.

**4. Xu hướng thời gian nằm viện theo nhóm tuổi:**

* Các nhóm tuổi trẻ nhất (0-9, 10-19, 20-29) có thời gian nằm viện trung bình thấp nhất và tương đối ổn định, dao động trong khoảng 6.56 đến 7.54 ngày. Điều này cho thấy đa số bệnh nhân ở độ tuổi này có khả năng phục hồi nhanh chóng.
* Các nhóm tuổi trung niên (40-49 và 50-59) có thời gian nằm viện trung bình tăng dần, lần lượt là 12.05 ngày và 10.45 ngày, cho thấy xu hướng nằm viện lâu hơn khi tuổi tăng lên.

**5. Ảnh hưởng của giá trị ngoại lệ:** Sự hiện diện của các giá trị ngoại lệ với thời gian nằm viện cực đoan, đặc biệt là ở nhóm 30-39 và 60+, có ảnh hưởng lớn đến việc tính toán và hiểu về thời gian nằm viện trung bình. Mặc dù chúng có thể là những trường hợp hiếm gặp, nhưng tác động của chúng đối với giá trị trung bình là rất đáng kể.

**Kết luận và Hướng nghiên cứu tiếp theo:** Phân tích này cho thấy thời gian nằm viện không chỉ đơn thuần tăng theo tuổi tác mà còn có những điểm bất thường đáng chú ý. Trường hợp của nhóm 30-39 với LOS trung bình cao nhất do các outliers cực đoan cần được điều tra sâu hơn để hiểu rõ nguyên nhân (ví dụ: các bệnh lý đặc biệt, tai nạn nghiêm trọng, hoặc vấn đề dữ liệu). Đối với nhóm 60+, xu hướng thời gian nằm viện dài hơn (mặc dù không phải cao nhất) và sự hiện diện của nhiều outliers vẫn khẳng định tầm quan trọng của việc quản lý bệnh lý nền và khả năng phục hồi chậm ở đối tượng này. Việc hiểu rõ các yếu tố này là then chốt để cải thiện chất lượng chăm sóc và tối ưu hóa thời gian nằm viện.

## 2.4.2. Theo giới tính:

### **2.4.2.1. Theo giới tính của R**

Hình 10: Phân bố thời gian nằm viện theo giới tính" - Histogram song song

Hình 11: "Phân bố thời gian nằm viện theo giới tính" - Boxplot với điểm trung bình

**Nhận xét chung:** Cặp biểu đồ này cung cấp cái nhìn chi tiết về sự phân bố và các đặc điểm thống kê (trung vị, trung bình, outliers) của thời gian nằm viện giữa Nam và Nữ.

**Hình 11 ("Phân bố thời gian nằm viện theo giới tính" - Histogram song song):**

**- Điểm nổi bật:** Cả Nam và Nữ đều có sự tập trung rất cao về số lượng bệnh nhân ở thời gian nằm viện rất ngắn (gần trục Y, từ 0 đến khoảng 10-20 ngày).

**- Sự tương đồng:** Nhìn chung, hình dạng phân bố của cả hai giới rất giống nhau, đều dốc mạnh về phía thời gian nằm viện ngắn và có một "đuôi dài" (long tail) về phía thời gian nằm viện dài hơn, mặc dù số lượng bệnh nhân giảm dần.

**- Điểm khác biệt nhỏ:** Quan sát kỹ, có vẻ như số lượng bệnh nhân Nữ tập trung ở thời gian nằm viện rất ngắn (gần 0) có phần cao hơn một chút so với Nam, nhưng sự khác biệt này không quá lớn khi nhìn vào tổng thể phân bố. Cả hai giới đều có những trường hợp nằm viện kéo dài, rải rác đến vài trăm ngày.

**- Tính trực quan:** Hiệu quả trong việc cho thấy hình dạng phân bố và sự tập trung dữ liệu. Dễ dàng so sánh hình thái phân bố giữa hai giới.

**Hình 10 ("Phân bố thời gian nằm viện theo giới tính" - Boxplot với điểm trung bình):**

**- Điểm nổi bật:**

**+ Trung vị (đường ngang trong hộp):** Gần như tương đồng giữa Nam và Nữ, cả hai đều rất thấp, cho thấy phần lớn bệnh nhân (cả Nam và Nữ) có thời gian nằm viện rất ngắn.

**+ Hộp (IQR):** Khoảng tứ phân vị của cả Nam và Nữ cũng rất hẹp và tương đồng, củng cố rằng 50% dữ liệu trung tâm của cả hai giới đều tập trung ở thời gian nằm viện ngắn.

**+ Điểm trung bình (chấm đỏ):** Vị trí của chấm đỏ (trung bình) cho cả Nam và Nữ gần như trùng khớp và nằm ngay trên đường trung vị, cho thấy phân bố có thể hơi lệch phải do các outliers.

**+ Outliers:** Cả hai giới đều có rất nhiều điểm ngoại lệ (chấm đen) trải dài lên trên, với một số trường hợp ở Nam đạt tới hơn 400 ngày, và ở Nữ đạt tới khoảng 100 ngày. Điều này xác nhận rằng cả hai giới đều có những trường hợp nằm viện rất dài, nhưng có vẻ Nam có outliers kéo dài hơn một chút (tuy nhiên, cần lưu ý đến số lượng và độ cực đoan của từng outlier).

**+ Tính trực quan:** Rất hiệu quả để so sánh nhanh chóng trung vị, khoảng phân vị và sự hiện diện của outliers giữa hai nhóm.

### **2.4.2.2. Theo giới tính của Python**

Hình 13: Boxplot thời gian nằm viện theo giới tính

Hình 12: "Phân phối thời gian nằm viện theo giới tính" (Kernel Density Estimate - KDE)

**Nhận xét chung:** Hai biểu đồ này cung cấp góc nhìn tương tự về phân bố và outliers theo giới tính, với biểu đồ mật độ hạt nhân cung cấp một cái nhìn "mượt mà" hơn về hình dạng phân phối.

**Hình 12 ("Boxplot thời gian nằm viện theo giới tính"):**

**- Điểm nổi bật:** Hoàn toàn khớp với nhận xét từ biểu đồ hộp của R. Trung vị (đường ngang trong hộp) và khoảng tứ phân vị (chiều cao hộp) của cả Nam và Nữ đều rất thấp và gần như giống hệt nhau, cho thấy phần lớn bệnh nhân ở cả hai giới đều nằm viện trong thời gian ngắn.

**- Outliers:** Cả Nam và Nữ đều có nhiều điểm ngoại lệ (vòng tròn xám) kéo dài lên trên. Nhóm Nam có một outlier cực đại vượt quá 550 ngày, tương tự như outlier đã thấy ở nhóm 30-39 trong phân tích trước đó (có thể outlier này là một bệnh nhân Nam ở độ tuổi 30-39). Nhóm Nữ cũng có nhiều outliers, nhưng chúng dường như không kéo dài đến mức cực đoan như outlier ở Nam.

**- Tính trực quan:** Cung cấp thông tin chi tiết và rõ ràng về phân bố, trung vị, và outliers.

**Hình 13 ("Phân phối thời gian nằm viện theo giới tính" - Histogram và KDE):**

**- Điểm nổi bật:** Củng cố nhận định về sự tương đồng trong phân bố thời gian nằm viện giữa Nam và Nữ. Cả hai đường cong phân phối (KDE) đều có đỉnh cao nhất ở gần 0 ngày và giảm dần nhanh chóng, sau đó có một "đuôi dài" về phía thời gian nằm viện dài hơn.

**- Sự tương đồng:** Các đường phân phối cho Nam (xanh ngọc) và Nữ (cam) gần như trùng khớp ở phần lớn các giá trị, đặc biệt là ở phần đỉnh và phần thân chính của phân bố (thời gian nằm viện ngắn).

**- Điểm khác biệt tinh tế:** Đường cong của Nữ có vẻ nhỉnh hơn một chút ở phần đỉnh (tức là có nhiều bệnh nhân Nữ hơn ở thời gian nằm viện rất ngắn), nhưng đây là một khác biệt rất nhỏ và có thể không có ý nghĩa thống kê lớn.

**- Tính trực quan:** Biểu đồ này rất tốt để hình dung hình dạng tổng thể của phân bố và so sánh mật độ của từng giới.

### **2.4.2.3. Tổng hợp nhận xét chung**

Dựa trên phân tích toàn diện của bốn biểu đồ về phân bố thời gian nằm viện (LOS) theo giới tính, có thể rút ra những nhận định quan trọng sau:

**1. Phân bố và Xu hướng chung:** Các biểu đồ đồng nhất cho thấy phần lớn bệnh nhân, bất kể giới tính, đều có thời gian nằm viện rất ngắn, tập trung mạnh mẽ ở khoảng 0 đến 10-20 ngày. Điều này được minh họa rõ ràng qua các đỉnh cao ở gần trục Y của biểu đồ histogram và KDE, cũng như các hộp (box) rất hẹp và thấp trong biểu đồ hộp.

**2. Sự tương đồng đáng kể giữa Nam và Nữ:** Một điểm nổi bật xuyên suốt các biểu đồ là sự tương đồng rất cao về phân bố thời gian nằm viện giữa Nam và Nữ.

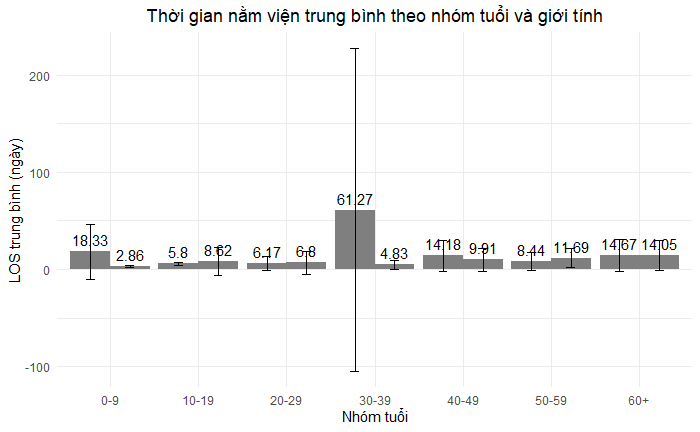
* Trung vị (median) của cả hai giới gần như trùng khớp và ở mức rất thấp, cho thấy một nửa số bệnh nhân của mỗi giới tính đều được xuất viện nhanh chóng.
* Khoảng tứ phân vị (IQR) cũng rất tương tự, chỉ ra rằng 50% dữ liệu trung tâm của cả Nam và Nữ có độ trải rộng tương đồng và tập trung ở các giá trị thấp.
* Hình dạng tổng thể của đường phân phối (histogram và KDE) cho Nam và Nữ gần như giống hệt nhau, đều có một đỉnh nhọn ở thời gian nằm viện ngắn và một đuôi dài về phía các giá trị cao hơn.

**3. Sự hiện diện của các trường hợp ngoại lệ (Outliers):** Mặc dù phần lớn bệnh nhân có LOS ngắn, cả hai giới tính đều có sự hiện diện của các trường hợp ngoại lệ với thời gian nằm viện kéo dài bất thường. Các biểu đồ hộp (cả từ R và Python) hiển thị rõ ràng các điểm outliers này. Đặc biệt, có một trường hợp outlier cực đoan ở bệnh nhân Nam với thời gian nằm viện vượt quá 500 ngày. Những outliers này, mặc dù hiếm, nhưng có thể ảnh hưởng đáng kể đến giá trị trung bình nếu chỉ dựa vào đó để so sánh.

**4. Không có sự khác biệt đáng kể về thời gian nằm viện trung bình giữa hai giới:** Do sự tương đồng về phân bố và vị trí của trung vị, có thể suy đoán rằng thời gian nằm viện trung bình của Nam và Nữ cũng sẽ rất gần nhau. Các điểm trung bình trong biểu đồ hộp của R cũng cho thấy sự trùng khớp của chúng. Các outliers, mặc dù kéo dài, nhưng do số lượng ít nên có thể không làm thay đổi đáng kể tổng thể phân bố giữa hai giới.

**Kết luận và Hướng nghiên cứu tiếp theo:** Nhìn chung, phân tích dữ liệu cho thấy giới tính không phải là yếu tố chính tạo ra sự khác biệt đáng kể về thời gian nằm viện ở mức độ tổng thể. Cả bệnh nhân Nam và Nữ đều có xu hướng nằm viện trong thời gian ngắn. Tuy nhiên, sự tồn tại của các trường hợp ngoại lệ với thời gian nằm viện cực đoan ở cả hai giới tính (đặc biệt là ở Nam với một outlier rất lớn) là một điểm cần được quan tâm. Để có cái nhìn sâu sắc hơn, nghiên cứu tiếp theo nên tập trung vào việc xác định nguyên nhân đằng sau các trường hợp nằm viện kéo dài này, có thể liên quan đến các bệnh lý cụ thể, tình trạng sức khỏe nền, hoặc các yếu tố xã hội, mà không chỉ dựa vào giới tính.

## 2.4.3. Theo giới tính và theo nhóm tuổi:



Hình 14: Thời gian nằm viện trung bình theo nhóm tuổi và giới tính

Biểu đồ này là một biểu đồ cột thể hiện thời gian nằm viện trung bình (LOS trung bình) được phân tách theo từng nhóm tuổi và trong mỗi nhóm tuổi lại được chia theo giới tính (mỗi cặp cột cạnh nhau trong một nhóm tuổi đại diện cho một giới tính, có lẽ cột đầu tiên là Nam và cột thứ hai là Nữ, hoặc ngược lại, nhưng vì chúng rất gần nhau nên ta có thể nhận định sự tương đồng).

**1. Điểm nổi bật nhất - Nhóm tuổi 30-39:**

* Thời gian nằm viện trung bình cực kỳ cao: Nhóm tuổi 30-39 tiếp tục cho thấy thời gian nằm viện trung bình cao nhất, lên tới 61.27 ngày. Đây là một con số vượt trội hoàn toàn so với tất cả các nhóm tuổi và giới tính khác.
* Biến động cực lớn: Cột lỗi (error bar) của nhóm 30-39 là khổng lồ, trải dài từ khoảng -100 đến hơn 200. Điều này không chỉ củng cố nhận định trước đó về sự hiện diện của các giá trị ngoại lệ cực đoan mà còn cho thấy rằng, trong nhóm tuổi này, có sự khác biệt rất lớn về thời gian nằm viện giữa các cá thể. Điều này có thể do một hoặc một vài trường hợp nằm viện kéo dài bất thường đã kéo giá trị trung bình lên rất cao**.**

**2. Nhóm tuổi 60+:**

* Thời gian nằm viện trung bình của nhóm 60+ là khoảng 14.05 - 14.67 ngày (hai cột nhỏ cạnh nhau). Con số này ổn định hơn và ít biến động hơn so với nhóm 30-39.
* Mặc dù không phải là cao nhất, nhưng vẫn cao hơn đáng kể so với các nhóm tuổi trẻ (0-29 tuổi). Điều này phù hợp với nhận định rằng người cao tuổi thường có bệnh lý đi kèm và thời gian phục hồi lâu hơn.

**3. Các nhóm tuổi trẻ hơn (0-9, 10-19, 20-29):**

* Thời gian nằm viện trung bình rất thấp và tương đối ổn định, dao động trong khoảng 2.86 - 7.5 ngày. Điều này cho thấy đa số bệnh nhân ở các nhóm tuổi này có thời gian nằm viện ngắn, khả năng phục hồi nhanh.
* Cột lỗi ở các nhóm này cũng rất nhỏ, chứng tỏ dữ liệu ít biến động hơn.

**4.** Các nhóm tuổi trung niên (40-49, 50-59):

Thời gian nằm viện trung bình tăng dần so với các nhóm trẻ, khoảng 9.91 - 14.18 ngày. Điều này phù hợp với xu hướng chung là thời gian nằm viện có thể tăng lên khi tuổi tác cao hơn.

**5. Ảnh hưởng của giới tính (trong từng nhóm tuổi):**

Trong từng cặp cột đại diện cho Nam và Nữ trong cùng một nhóm tuổi, giá trị trung bình của hai giới rất gần nhau. Điều này gợi ý rằng giới tính không phải là yếu tố chính tạo ra sự khác biệt lớn về thời gian nằm viện trung bình trong cùng một nhóm tuổi. Hay nói cách khác, thời gian nằm viện trung bình chủ yếu bị ảnh hưởng bởi nhóm tuổi, chứ không phải giới tính.

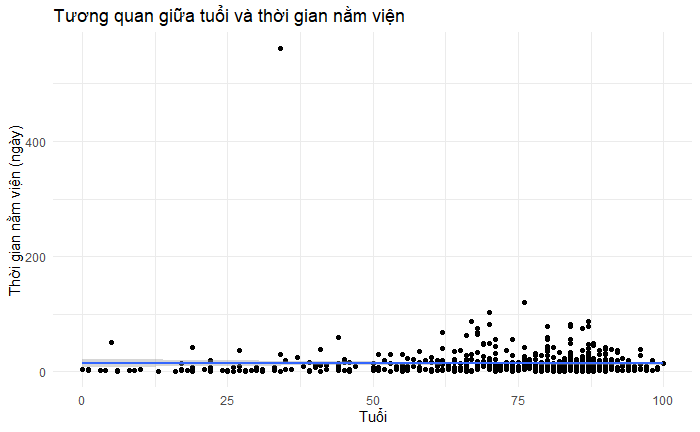
**Kết luận tổng hợp:**

Biểu đồ này khẳng định lại rằng nhóm tuổi 30-39 có thời gian nằm viện trung bình cao nhất một cách rõ rệt (61.27 ngày), mặc dù có sự biến động rất lớn trong nhóm này. Các nhóm tuổi trẻ có thời gian nằm viện ngắn nhất, và thời gian nằm viện có xu hướng tăng lên ở các nhóm tuổi lớn hơn (từ 40-49 trở lên). Quan trọng hơn, biểu đồ này cho thấy giới tính dường như không phải là một yếu tố ảnh hưởng đáng kể đến thời gian nằm viện trung bình khi xem xét trong từng nhóm tuổi cụ thể, vì thời gian nằm viện trung bình giữa Nam và Nữ trong cùng một nhóm tuổi là rất tương đồng.

Hướng nghiên cứu tiếp theo: Trường hợp đặc biệt của nhóm 30-39 với LOS trung bình cao và biến động lớn cần được điều tra sâu hơn. Có thể có một nguyên nhân đặc thù nào đó (ví dụ: một loại bệnh lý nghiêm trọng phổ biến ở độ tuổi này, tai nạn nặng, hoặc có thể là lỗi dữ liệu) đã làm tăng LOS của nhóm này lên đáng kể

# Phân tích đa biến

## Tương quan giữa tuổi và thời gian nằm viện

Nhận xét chung:

Hình 15: Tương quan giữa tuổi và thời gian nằm viện

### **2.5.1.1. Phân tán dữ liệu:**

Phần lớn các điểm dữ liệu (tượng trưng cho từng bệnh nhân) tập trung ở phía dưới của biểu đồ, đặc biệt là ở những giá trị thời gian nằm viện thấp (dưới 50 ngày). Điều này cho thấy đa số bệnh nhân, ở mọi lứa tuổi, có thời gian nằm viện ngắn.

Càng về phía tuổi cao hơn (ví dụ từ khoảng 60 tuổi trở lên), các điểm dữ liệu dường như phân tán rộng hơn về phía trên, cho thấy có nhiều bệnh nhân lớn tuổi hơn có thời gian nằm viện kéo dài.

### **2.5.1.2. Xu hướng tương quan:**

Đường hồi quy (đường màu xanh dương) có xu hướng đi lên nhẹ. Điều này gợi ý rằng có một tương quan dương yếu giữa tuổi và thời gian nằm viện. Nói cách khác, nhìn chung, khi tuổi tác tăng lên, thời gian nằm viện có xu hướng tăng nhẹ.

Tuy nhiên, đường hồi quy này tương đối phẳng, cho thấy mối quan hệ này không mạnh mẽ hoặc không tuyến tính rõ rệt trên toàn bộ phạm vi tuổi.

### **2.5.1.3. Sự hiện diện của các giá trị ngoại lệ và biến động lớn:**

Có những điểm dữ liệu nằm rất xa so với phần lớn các điểm khác, đặc biệt là một điểm nằm viện gần 500 ngày ở độ tuổi khoảng 30. Điểm này phù hợp với các nhận định trước đây về outlier cực đoan ở nhóm 30-39 tuổi đã được nhìn thấy trong các biểu đồ hộp và biểu đồ trung bình.

Ngoài ra, có những điểm nằm viện dài hơn (ví dụ từ 50-100 ngày) rải rác ở nhiều độ tuổi khác nhau, nhưng tập trung rõ rệt hơn ở các độ tuổi từ 60 trở lên.

Biến động (độ phân tán của các điểm quanh đường hồi quy) dường như lớn hơn ở các nhóm tuổi cao hơn. Điều này có nghĩa là ở người lớn tuổi, thời gian nằm viện có thể rất ngắn hoặc rất dài, cho thấy sự đa dạng lớn trong tình trạng sức khỏe và quá trình phục hồi**.**

### **2.5.1.4. Kết luận về tương quan:**

Mặc dù có xu hướng tăng nhẹ về thời gian nằm viện khi tuổi tăng, mối quan hệ này không chặt chẽ. Có nghĩa là, biết tuổi của một người không đủ để dự đoán chính xác thời gian nằm viện của họ. Các yếu tố khác ngoài tuổi (như bệnh lý nền, mức độ nghiêm trọng của bệnh, điều kiện sức khỏe tổng thể, v.v.) có thể đóng vai trò quan trọng hơn trong việc quyết định thời gian nằm viện thực tế, đặc biệt là ở những trường hợp nằm viện kéo dài.

## **Phân tích ANOVA theo giới tính**:

1. Df Sum Sq Mean Sq F value Pr(>F)
2. factor(Gender) 1 1278 1277.6 1.912 0.167
3. Residuals 674 450355 668.2

Nhận xét và kết luận từ phân tích ANOVA:

1. **Giả thuyết không (H0):** Không có sự khác biệt có ý nghĩa thống kê về thời gian nằm viện trung bình giữa Nam và Nữ.
2. **Giá trị P (p-value):** Giá trị p là 0.167.
3. **Mức ý nghĩa (alpha level):** Thông thường, chúng ta sử dụng mức ý nghĩa α=0.05.
4. **Kết luận:** Vì giá trị p (0.167) **lớn hơn** mức ý nghĩa α (0.05), chúng ta **không có đủ bằng chứng thống kê để bác bỏ giả thuyết không**.

**Kết luận:**

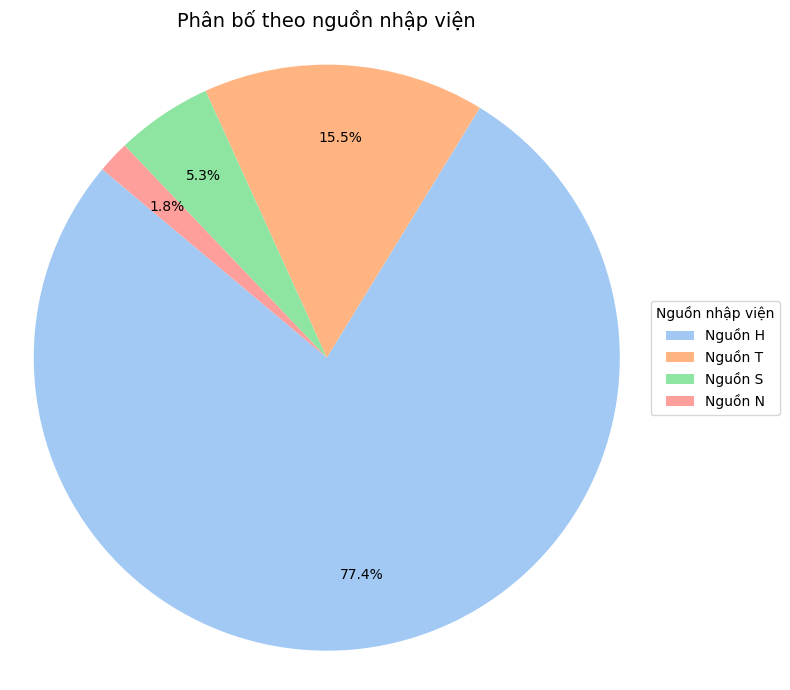
Phân tích ANOVA này cho thấy **không có sự khác biệt có ý nghĩa thống kê về thời gian nằm viện trung bình giữa giới tính Nam và Nữ** trong tập dữ liệu này. Điều này củng cố các nhận định trước đó từ các biểu đồ hộp và biểu đồ phân phối, nơi mà đường trung vị, các hộp và hình dạng phân phối của thời gian nằm viện giữa hai giới tính Nam và Nữ là rất tương đồng.

Mặc dù có những điểm ngoại lệ (outliers) khác nhau giữa Nam và Nữ (như outlier cực đoan ở Nam), nhưng về mặt thống kê, khi xem xét tổng thể phương sai, giới tính không phải là yếu tố giải thích sự khác biệt đáng kể trong thời gian nằm viện trung bình.

# Phân tích theo nguồn nhập viện

Bảng 5: Bảng phân bố theo nguồn nhập viện

|  |  |  |
| --- | --- | --- |
| **Nguồn nhập viện** | **Số lượng** | **Tỷ lệ (%)** |
| Nguồn H | 523 | 77.37 |
| Nguồn T | 105 | 15.53 |
| Nguồn S | 36 | 5.33 |
| Nguồn N | 12 | 1.78 |

****

Hình 16: Biểu đồ tròn phân bố theo nguồn nhập viện

Nhận xét về Phân bố theo nguồn nhập viện:

Biểu đồ tròn và bảng số liệu đi kèm cung cấp cái nhìn rõ ràng về các kênh bệnh nhân được đưa đến bệnh viện. Việc hiểu rõ ý nghĩa của từng mã nguồn nhập viện (Admission Source) theo chú thích đã cung cấp sẽ làm tăng giá trị của nhận xét.

## 2.6.1. Nguồn H (Recognised hospital - Bệnh viện được công nhận) là kênh nhập viện chủ yếu và quan trọng nhất:

* **Chiếm tỷ lệ áp đảo 77.37% (tương ứng 523 bệnh nhân)** trong tổng số các trường hợp nhập viện.
* Theo chú thích, "Nguồn H" là khi bệnh nhân được **chuyển đến từ một bệnh viện khác.** Điều này nhấn mạnh vai trò của bệnh viện đang được phân tích như một cơ sở tuyến trên hoặc một trung tâm chuyên khoa, tiếp nhận các trường hợp bệnh nhân đã được thăm khám hoặc điều trị ban đầu tại các cơ sở y tế khác. Sự phổ biến của nguồn này cho thấy một hệ thống chuyển tuyến hoặc liên kết giữa các bệnh viện hoạt động mạnh mẽ, đảm bảo bệnh nhân được chuyển đến nơi phù hợp với tình trạng bệnh của họ.

## 2.6.2. Nguồn T (Other - Nguồn khác) là kênh nhập viện thứ cấp đáng kể:

* **Chiếm 15.53% (105 bệnh nhân) các trường hợp nhập viện.**
* Theo chú thích, "Nguồn T" bao gồm những nơi không thuộc các danh mục khác, ví dụ: nơi làm việc, đường phố, bãi biển, v.v. Điều này gợi ý rằng đây là kênh tiếp nhận các trường hợp khẩn cấp, tai nạn hoặc bệnh đột xuất từ cộng đồng, không thông qua một cơ sở y tế ban đầu nào. Vai trò của bệnh viện trong việc xử lý các tình huống cấp cứu trực tiếp từ bên ngoài là rõ ràng.

## 2.6.3. Nguồn S (Other health care establishment - Cơ sở chăm sóc sức khỏe khác) và Nguồn N (Nursing home or other residential aged care - Viện dưỡng lão/cơ sở chăm sóc người cao tuổi nội trú khác) chiếm tỷ lệ rất nhỏ:

* **"Nguồn S" chiếm 5.33% (36 bệnh nhân).** Đây là những bệnh nhân được chuyển đến từ các cơ sở chăm sóc sức khỏe ban đầu như trung tâm y tế, phòng khám chuyên khoa, hoặc phòng khám bác sĩ đa khoa (GP). Tỷ lệ thấp này có thể chỉ ra rằng các trường hợp cần nhập viện từ các phòng khám này không nhiều, hoặc bệnh viện không phải là điểm đến chính cho các ca bệnh thông thường được giới thiệu từ cấp cơ sở.
* **"Nguồn N" là kênh ít phổ biến nhất, chỉ với 1.78% (12 bệnh nhân).** Đây là những bệnh nhân được chuyển đến từ viện dưỡng lão hoặc cơ sở chăm sóc người cao tuổi nội trú khác. Tỷ lệ rất thấp này có thể phản ánh rằng bệnh viện không phải là nơi tiếp nhận chính các trường hợp từ các cơ sở chăm sóc dài hạn này, hoặc các cơ sở đó có hệ thống chăm sóc riêng hoặc chuyển đến các bệnh viện chuyên biệt hơn.

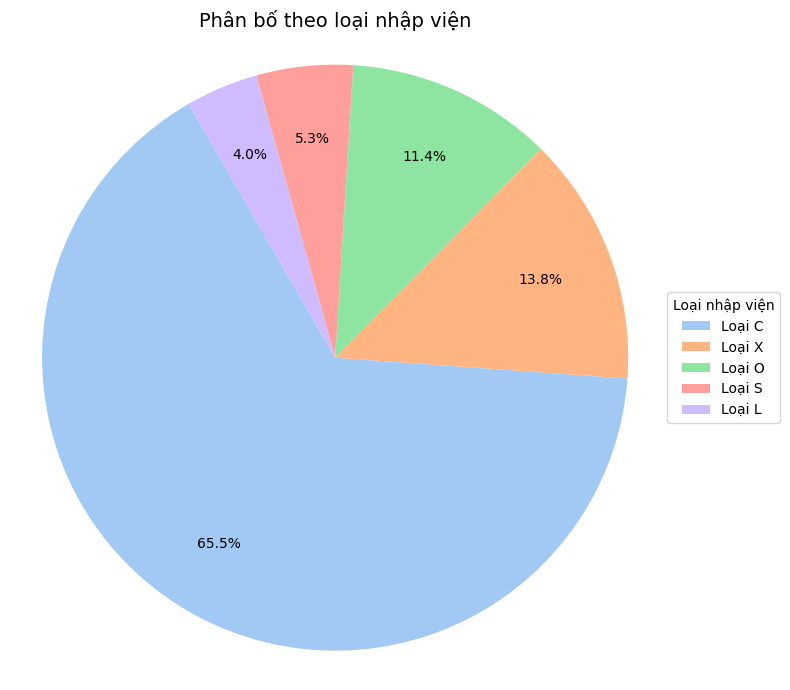
## 2.6.4. Kết luận tổng thể về Nguồn nhập viện:

Phân tích chi tiết về nguồn nhập viện cho thấy bệnh viện này chủ yếu hoạt động như một **cơ sở tiếp nhận tuyến trên hoặc trung tâm cấp cứu.** Đại đa số bệnh nhân được **chuyển đến từ các bệnh viện khác (Nguồn H)**, phản ánh vai trò của bệnh viện trong hệ thống y tế có tính phân cấp hoặc chuyên môn hóa. Kênh cấp cứu trực tiếp từ cộng đồng (Nguồn T) cũng là một phần quan trọng. Ngược lại, bệnh viện ít tiếp nhận bệnh nhân từ các cơ sở chăm sóc ban đầu hoặc viện dưỡng lão, cho thấy có thể đây không phải là chức năng chính của bệnh viện hoặc có các kênh chuyển tiếp khác hiệu quả hơn cho những trường hợp đó. Việc hiểu rõ nguồn nhập viện giúp bệnh viện tối ưu hóa quy trình tiếp nhận, phân bổ nguồn lực và phát triển các chiến lược hợp tác với các cơ sở y tế khác.

# Phân bố bệnh nhân theo nhóm tuổi

Bảng 6: Bảng phân bố theo loại nhập viện

|  |  |  |
| --- | --- | --- |
| **Loại nhập viện** | **Số lượng** | **Tỷ lệ (%)** |
| Loại C | 443 | 65.53 |
| Loại X | 93 | 13.76 |
| Loại O | 77 | 11.39 |
| Loại S | 36 | 5.33 |
| Loại L | 27 | 3.99 |



Hình 17: Biểu đồ tròn phân bố theo loại nhập viện

**Nhận xét về Phân bố bệnh nhân theo Loại nhập viện (Admission Type)**

Biểu đồ tròn và bảng số liệu đi kèm hiển thị rõ ràng các loại hình nhập viện mà bệnh viện tiếp nhận. Việc hiểu rõ ý nghĩa của từng mã loại nhập viện theo chú thích đã cung cấp sẽ giúp chúng ta đưa ra những nhận định có giá trị.

## 2.7.1. Loại C (Newborn Admission - Trẻ sơ sinh nhập viện) là loại hình nhập viện chủ yếu:

* **Chiếm tỷ lệ áp đảo với 65.53% (tương ứng 443 bệnh nhân) tổng số trường hợp nhập viện.**
* Theo chú thích, "Loại C" được sử dụng riêng cho **trẻ sơ sinh được sinh ra tại bệnh viện và cần được chăm sóc.** Điều này cho thấy bệnh viện có chức năng sản khoa hoặc khoa nhi sơ sinh rất mạnh và là nơi sinh chủ yếu. Hơn 2/3 số lượt nhập viện là dành cho trẻ sơ sinh, đây là một điểm đặc trưng và nổi bật của bệnh viện này.

## 2.7.2. Loại X (Dữ liệu cũ) và Loại O (Posthumous Organ Procurement - Tiếp nhận để lấy tạng sau khi chết) chiếm tỷ lệ đáng kể:

* **"Loại X" chiếm 13.76% (93 bệnh nhân).** Chú thích cho biết đây là "dữ liệu cũ". Tỷ lệ này là khá cao đối với một loại dữ liệu được dán nhãn "cũ". Điều này có thể ám chỉ những trường hợp nhập viện được ghi nhận theo một hệ thống hoặc quy chuẩn cũ, không còn áp dụng. Cần làm rõ thêm về ý nghĩa của "dữ liệu cũ" để hiểu đúng bản chất của các lượt nhập viện này.
* **"Loại O" chiếm 11.39% (77 bệnh nhân).** Đây là những trường hợp rất đặc biệt, liên quan đến **tiếp nhận để lấy tạng sau khi chết.** Tỷ lệ này (hơn 1/10 tổng số lượt nhập viện) là **cực kỳ cao và bất thường** đối với loại hình này. Trong thực tế lâm sàng, các trường hợp hiến tạng sau khi chết thường rất hiếm. Việc "Loại O" chiếm tỷ lệ cao đến vậy có thể là một **điểm dữ liệu cần được kiểm tra kỹ lưỡng (potential data anomaly)**, hoặc bệnh viện này là một trung tâm chuyên biệt về cấy ghép tạng lớn, thu hút số lượng lớn các trường hợp hiến tạng.

## 2.7.3. Các loại S (Hospital Transfer - Chuyển viện) và L (Statistical Admission - Nhập viện thống kê) chiếm tỷ lệ nhỏ:

* **"Loại S" chiếm 5.33% (36 bệnh nhân).** Đây là các bệnh nhân được **chuyển trực tiếp từ một bệnh viện khác.** So với "Nguồn H" (Recognised hospital) ở phân tích nguồn nhập viện (77.37%), "Loại S" (Hospital Transfer) lại có tỷ lệ thấp hơn nhiều. Điều này có thể ngụ ý rằng không phải tất cả các trường hợp chuyển viện từ bệnh viện khác đều được phân loại là "Hospital Transfer" (Loại S) trong "Admission Type", hoặc "Loại S" chỉ áp dụng cho một số loại chuyển viện cụ thể, trong khi các trường hợp chuyển viện thông thường khác có thể được xếp vào "Loại M" (Other Admission - Nhập viện thông thường).
* **"Loại L" chiếm 3.99% (27 bệnh nhân).** Đây là "nhập viện thống kê", không phải nhập viện thực sự từ bên ngoài, mà dùng cho **mục đích hành chính khi bệnh nhân thay đổi loại hình chăm sóc trong cùng bệnh viện.** Tỷ lệ này cho thấy có một số lượng nhỏ các trường hợp thay đổi loại hình chăm sóc nội viện được ghi nhận như một lượt nhập viện mới.

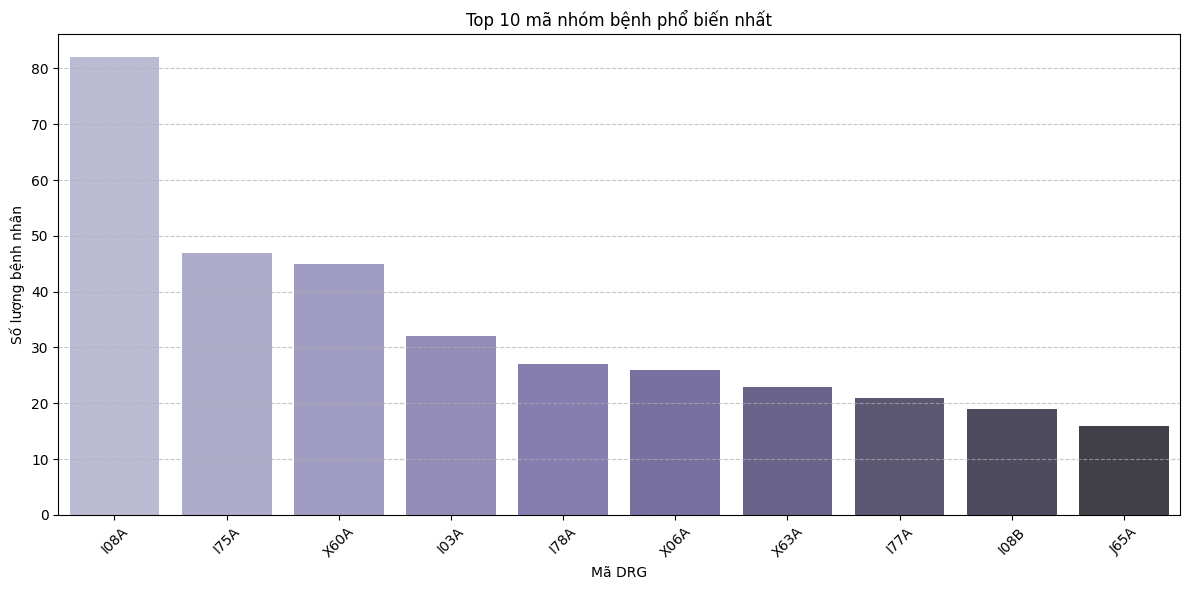
## 2.7.4. Thiếu loại M (Other Admission - Nhập viện thông thường):

* Điều đáng chú ý là "Loại M" (Nhập viện thông thường), được chú thích là loại nhập viện phổ biến nhất cho các trường hợp từ cộng đồng, danh sách chờ, hoặc phòng khám ngoại trú, lại **không xuất hiện** trong biểu đồ và bảng số liệu này. Điều này có thể là do:
  + Dữ liệu đã được lọc bỏ "Loại M" khi tạo biểu đồ.
  + Tất cả các trường hợp nhập viện thông thường đã được phân loại chi tiết hơn vào các nhóm khác (như "Loại C", "Loại X", v.v.) dựa trên các tiêu chí khác, khiến "Loại M" trở nên không đáng kể hoặc không được ghi nhận trong tập dữ liệu cụ thể này. Đây là một điểm cần làm rõ để hiểu bức tranh toàn diện về các lượt nhập viện.

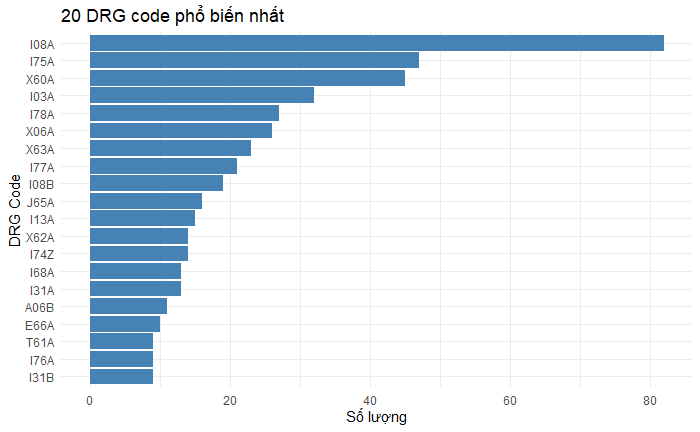
## 2.7.5. Kết luận tổng thể về Loại nhập viện:

Phân tích loại nhập viện cho thấy bệnh viện này có một chức năng sản/nhi khoa rất mạnh mẽ, với **hơn 2/3 số lượt nhập viện là trẻ sơ sinh (Loại C)**. Sự hiện diện đáng kể của **"Loại X" (dữ liệu cũ)** và đặc biệt là **"Loại O" (tiếp nhận để lấy tạng sau khi chết) với tỷ lệ bất thường cao** là những điểm cần được điều tra và xác minh kỹ lưỡng về tính chính xác của dữ liệu hoặc ý nghĩa thực tế. Các loại hình nhập viện khác như chuyển viện (Loại S) và nhập viện thống kê (Loại L) chiếm tỷ lệ nhỏ hơn. Việc thiếu vắng "Loại M" (Nhập viện thông thường) cũng là một câu hỏi cần giải đáp để có cái nhìn đầy đủ về các nguồn và loại hình bệnh nhân mà bệnh viện này tiếp nhận.

# Phân tích theo DRG code ( mã nhóm bệnh)



Hình 18: Top 10 mã nhóm bệnh phổ biến



Hình 19: Top 20 mã nhóm bệnh phổ biến

**Nhận xét về Phân bố các mã DRG phổ biến nhất:**

Hai biểu đồ này (biểu đồ thanh ngang và biểu đồ cột dọc) đều hiển thị các mã DRG phổ biến nhất trong bộ dữ liệu, cho phép chúng ta nhận diện các nhóm bệnh chính mà bệnh viện đang điều trị.

## 2.8.1. Mã I08A - Bệnh về tim mạch (Cardiovascular Diseases) là nhóm bệnh phổ biến nhất:

* **Mã I08A đứng đầu về số lượng, với hơn 80 trường hợp.** Mặc dù không có mô tả trực tiếp về mã I08A trong chú thích DRG, nhưng với ký hiệu "I", đây thường là các mã liên quan đến bệnh hệ tuần hoàn (theo chuẩn ICD-10). Điều này cho thấy bệnh viện này xử lý một số lượng lớn các trường hợp bệnh tim mạch, có thể là một chuyên khoa mạnh của bệnh viện hoặc phản ánh gánh nặng bệnh tật chung trong cộng đồng.
* Việc mã này vượt trội hơn hẳn các mã khác cho thấy đây là lĩnh vực trọng tâm trong hoạt động của bệnh viện.

## 2.8.2. Các mã phổ biến tiếp theo:

* **I75A và X60A là hai mã phổ biến tiếp theo**, với số lượng lần lượt là khoảng 47 và 45 trường hợp. Tương tự I08A, mã I75A cũng có thể liên quan đến bệnh tim mạch hoặc tuần hoàn.
* **I03A, I78A, X06A, X63A, I77A, I08B, J65A** cũng là các mã DRG có số lượng bệnh nhân đáng kể, tạo nên top 10 mã bệnh phổ biến nhất.
  + Các mã bắt đầu bằng "I" (như I03A, I78A, I77A, I08B) tiếp tục gợi ý về sự phổ biến của các bệnh liên quan đến hệ tuần hoàn.
  + Các mã bắt đầu bằng "X" (như X60A, X06A, X63A) thường liên quan đến các nguyên nhân bên ngoài gây bệnh hoặc tử vong (ví dụ: tai nạn, chấn thương, ngộ độc) theo phân loại ICD-10. Điều này cho thấy bệnh viện cũng tiếp nhận một lượng đáng kể các trường hợp chấn thương/ngoại khoa.
  + Mã "J65A" có thể liên quan đến các bệnh hô hấp (theo phân loại ICD-10).

## 2.8.3. Sự đa dạng của các mã DRG:

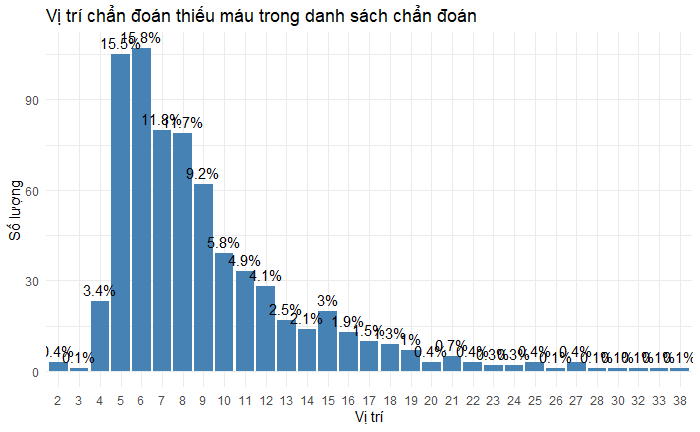
* Biểu đồ 20 mã DRG phổ biến nhất cho thấy sự đa dạng của các loại bệnh mà bệnh viện điều trị, mặc dù có một số mã nổi trội hơn hẳn. Các mã từ vị trí thứ 11 trở đi có số lượng bệnh nhân giảm dần và tương đối đồng đều.

## 2.8.4. Kết luận:

Phân tích các mã DRG phổ biến nhất cho thấy **bệnh viện này có chuyên môn hoặc tiếp nhận số lượng lớn các trường hợp bệnh lý về hệ tuần hoàn (tim mạch) và các trường hợp liên quan đến nguyên nhân bên ngoài (chấn thương, tai nạn).** Sự đa dạng của các mã DRG cũng phản ánh rằng bệnh viện điều trị nhiều loại bệnh khác nhau, không chỉ tập trung vào một vài chuyên khoa duy nhất. Việc xác định các mã DRG phổ biến này rất quan trọng để bệnh viện có thể phân bổ nguồn lực (nhân lực, trang thiết bị) một cách hiệu quả, tối ưu hóa quy trình điều trị và quản lý chi phí

# Phân tích vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán

Hình 20: Phân bố vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán của Python

****

Hình 21: Phân bố vị trí chẩn đoán thiếu máu trong danh sách của R

**Nhận xét về "Phân bố vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán" (Kết hợp số lượng và tỷ lệ %)**

Hai biểu đồ này (một thể hiện số lượng và một thể hiện tỷ lệ phần trăm) cùng với bảng số liệu chi tiết, cung cấp thông tin quan trọng về vị trí mà chẩn đoán thiếu máu xuất hiện trong danh sách chẩn đoán của bệnh nhân. Chú thích "1 = chính, 2+ = phụ" là mấu chốt để hiểu bản chất của các vị trí này.

## 2.9.1. Thiếu máu hiếm khi là chẩn đoán chính:

Chỉ có 3 bệnh nhân (chiếm 0.4%) có chẩn đoán thiếu máu ở vị trí 1 (tình trạng chính). Điều này cực kỳ quan trọng, cho thấy thiếu máu rất ít khi là lý do chính khiến bệnh nhân nhập viện.

## 2.9.2. Thiếu máu thường xuyên là chẩn đoán phụ, đặc biệt ở các vị trí 5, 6 và 7:

* Số lượng bệnh nhân có chẩn đoán thiếu máu tăng vọt đáng kể ở các vị trí phụ (từ 2 trở đi).
* Vị trí 5 có số lượng cao nhất với 105 trường hợp, chiếm 15.5% tổng số chẩn đoán thiếu máu.
* Vị trí 6 đứng thứ hai với 107 trường hợp, chiếm 15.8%.
* Vị trí 7 theo sau với 80 trường hợp, chiếm 11.8%.
* Điều này chỉ ra rằng thiếu máu chủ yếu là một tình trạng liên quan (Associated condition) hoặc tình trạng biến chứng (Complicating condition) của bệnh lý chính khác. Sự tập trung đặc biệt ở các vị trí 5, 6, 7 cho thấy nó thường không phải là chẩn đoán phụ đầu tiên, mà là một trong những chẩn đoán phụ xuất hiện sau trong danh sách.

## 2.9.3. Số lượng chẩn đoán thiếu máu giảm dần ở các vị trí phụ tiếp theo:

* Sau vị trí 7, số lượng và tỷ lệ bệnh nhân có chẩn đoán thiếu máu giảm dần đều đặn. Mặc dù vẫn có những trường hợp thiếu máu được ghi nhận ở các vị trí rất xa (lên đến vị trí 38), nhưng số lượng là rất nhỏ (chỉ 1-3 bệnh nhân cho mỗi vị trí).
* Điều này củng cố nhận định rằng nếu thiếu máu xuất hiện như một chẩn đoán phụ, nó có xu hướng được ghi nhận ở các vị trí tương đối sớm trong danh sách chẩn đoán phụ.

## 2.9.4. Ý nghĩa lâm sàng và quản lý:

* Thiếu máu là một dấu hiệu hoặc biến chứng, không phải nguyên nhân chính: Dữ liệu này mạnh mẽ gợi ý rằng thiếu máu thường là một hậu quả hoặc một phần của bức tranh bệnh lý phức tạp hơn, chứ không phải là lý do nhập viện ban đầu. Điều này có ý nghĩa quan trọng trong việc chẩn đoán và điều trị, vì các bác sĩ cần tìm kiếm và điều trị nguyên nhân gốc rễ gây ra thiếu máu thay vì chỉ tập trung vào triệu chứng.
* Tầm quan trọng của việc ghi nhận đầy đủ chẩn đoán phụ: Việc thiếu máu xuất hiện ở các vị trí phụ cho thấy tầm quan trọng của việc ghi nhận nhiều chẩn đoán cho một bệnh nhân để có cái nhìn toàn diện về tình trạng sức khỏe của họ.

## 2.9.5. Kết luận:

Phân tích "Phân bố vị trí chẩn đoán thiếu máu trong danh sách chẩn đoán" khẳng định rằng thiếu máu gần như luôn là một tình trạng phụ (liên quan hoặc biến chứng), với tỷ lệ xuất hiện cao nhất ở các vị trí chẩn đoán 5, 6 và 7. Điều này cung cấp một cái nhìn quan trọng về vai trò của thiếu máu trong bối cảnh bệnh lý của bệnh nhân, gợi ý rằng nó thường là một hệ quả của các bệnh chính khác cần được chú ý trong quá trình điều trị.

# III. PHÂN TÍCH NÂNG CAO

Sau khi đã tiến hành mô tả tổng quan đặc điểm nhân khẩu học và thời gian nằm viện, phần này tập trung khai thác sâu hơn các khía cạnh quan trọng có thể ảnh hưởng đến gánh nặng bệnh tật và quá trình điều trị bệnh nhân thiếu máu do thiếu sắt, thông qua các kỹ thuật phân tích nâng cao:

# 3.1. Phân tích ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng

## 3.1.1. Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R



Hình 22: Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R

**Nhận xét về "Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng"**

Biểu đồ ma trận tương quan này hiển thị mối quan hệ tuyến tính giữa ba biến: Tuổi (age\_years), LOS (thời gian nằm viện - số ngày), và Cân nặng (weight). Màu sắc và giá trị số trên biểu đồ thể hiện hướng và độ mạnh của mối tương quan.

**1. Mối tương quan giữa Tuổi và LOS (Thời gian nằm viện):**

* Hệ số tương quan giữa Tuổi và LOS là **0.08**.
* Đây là một **mối tương quan dương, nhưng rất yếu**. Điều này có nghĩa là có một xu hướng rất nhỏ: khi tuổi tăng, thời gian nằm viện có thể hơi tăng lên, nhưng mối liên hệ này không đáng kể. Nó cho thấy tuổi tác không phải là yếu tố dự đoán mạnh mẽ cho thời gian nằm viện trong tập dữ liệu này.

**2. Mối tương quan giữa Tuổi và Cân nặng:**

* Hệ số tương quan giữa Tuổi và Cân nặng là **-0.08**.
* Đây là một **mối tương quan âm, cũng rất yếu**. Điều này chỉ ra rằng có một xu hướng rất nhỏ: khi tuổi tăng, cân nặng có thể hơi giảm đi, hoặc ngược lại, nhưng mối liên hệ này cũng không có ý nghĩa đáng kể.

**3. Mối tương quan giữa LOS (Thời gian nằm viện) và Cân nặng:**

* Hệ số tương quan giữa LOS và Cân nặng là **-0.03**.
* Đây là một **mối tương quan âm, cực kỳ yếu, gần như bằng 0**. Điều này cho thấy hầu như không có mối quan hệ tuyến tính giữa thời gian nằm viện và cân nặng trong dữ liệu này. Cân nặng của bệnh nhân không có xu hướng liên quan đến việc họ nằm viện ngắn hay dài ngày.

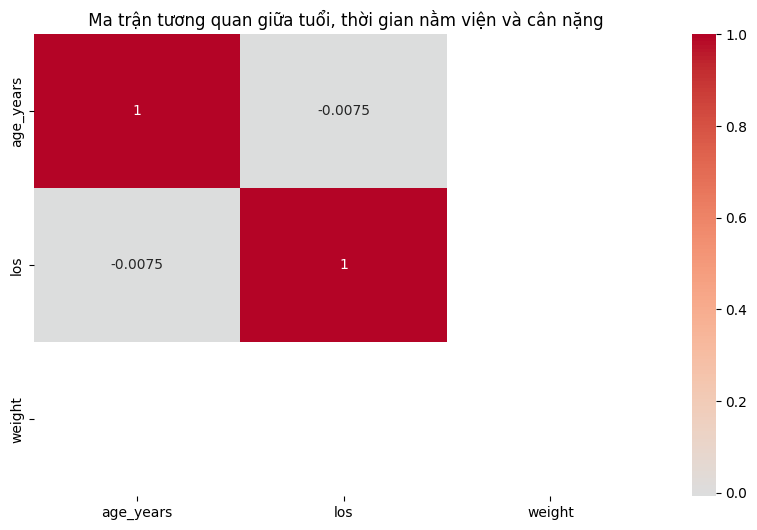
**Kết luận tổng thể:**

Nhìn chung, biểu đồ ma trận tương quan cho thấy **mối quan hệ tuyến tính giữa các cặp biến Tuổi, LOS và Cân nặng là rất yếu hoặc không đáng kể.** Tất cả các hệ số tương quan đều rất gần với 0 (0.08, -0.08, -0.03).

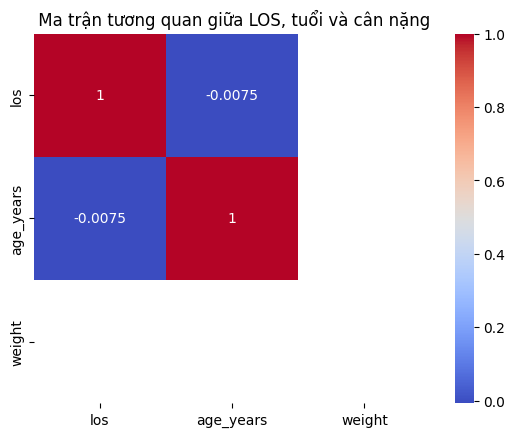
**Hàm ý:**

* Các biến này (Tuổi, LOS, Cân nặng) dường như **không có mối quan hệ tuyến tính mạnh mẽ** với nhau trong tập dữ liệu được phân tích.
* Điều này ngụ ý rằng các biến này, khi xét riêng lẻ và theo mối quan hệ tuyến tính đơn thuần, không phải là yếu tố dự đoán chính cho nhau.
* Để hiểu rõ hơn về các yếu tố ảnh hưởng đến LOS, Tuổi hay Cân nặng, có thể cần xem xét các mối quan hệ phi tuyến tính, hoặc các biến khác trong bộ dữ liệu (ví dụ: loại bệnh, loại hình chăm sóc, độ nặng của bệnh, tình trạng sức khỏe nền, v.v.), hoặc phân tích tương quan cho các nhóm nhỏ hơn trong dữ liệu.

## 3.1.2. Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của Python



Hình 23: Ma trận tương quan giữa tuổi, thời gian nằm viện và cân nặng của R



Hình 24: Ma trận tương quan giữa LOS, tuổi và cân nặng

**Nhận xét chung về ảnh hưởng của Tuổi và Giới tính đến Thời gian nằm viện (LOS) qua Hồi quy OLS và ANOVA**

Hai phân tích này cung cấp cái nhìn về mối quan hệ giữa LOS với tuổi và giới tính.

**1. Kết quả Hồi quy OLS**

* **Mô hình tổng thể:**
  + **R-squared = 0.003:** Giá trị này cực kỳ thấp, cho thấy mô hình hồi quy chỉ giải thích được **0.3%** sự biến thiên của thời gian nằm viện (LOS). Điều này có nghĩa là tuổi và giới tính (dưới dạng biến Gender\_code) chỉ giải thích được một phần rất nhỏ, gần như không đáng kể, sự khác biệt về LOS giữa các bệnh nhân.
  + **Adj. R-squared = -0.000:** Giá trị R-squared điều chỉnh gần bằng 0 hoặc âm càng củng cố rằng mô hình này không phù hợp để giải thích LOS.
  + **F-statistic = 0.9569, Prob (F-statistic) = 0.385:** Giá trị p của F-statistic (0.385) lớn hơn mức ý nghĩa thông thường (0.05). Điều này cho thấy **mô hình tổng thể không có ý nghĩa thống kê**, tức là các biến age\_years và Gender\_code *cùng nhau* không giải thích đáng kể sự thay đổi của LOS.
* **Tác động của từng biến độc lập:**
  + **age\_years (Tuổi):**
    - coef = -0.0034: Hệ số này rất nhỏ và âm, cho thấy khi tuổi tăng 1 đơn vị, LOS trung bình giảm đi 0.0034 ngày, nhưng đây là một mức thay đổi không đáng kể.
    - P>|t| = 0.946: Giá trị p này rất lớn (gần 1), cao hơn nhiều so với 0.05. Điều này có nghĩa là **biến age\_years không có tác động có ý nghĩa thống kê đến LOS** khi các yếu tố khác được giữ nguyên trong mô hình này.
  + **Gender\_code (Giới tính):**
    - coef = -2.8419: Hệ số này cho thấy có sự khác biệt về LOS trung bình giữa hai nhóm giới tính (tùy thuộc vào cách mã hóa giới tính, ví dụ 0 cho Nam, 1 cho Nữ, thì Nữ có LOS thấp hơn Nam 2.84 ngày).
    - P>|t| = 0.171: Giá trị p này lớn hơn 0.05. Điều này cho thấy **biến Gender\_code cũng không có tác động có ý nghĩa thống kê đến LOS** trong mô hình hồi quy này.
* **Kết luận từ OLS:** Mô hình hồi quy này cho thấy **tuổi và giới tính không phải là các yếu tố dự báo có ý nghĩa thống kê đối với thời gian nằm viện (LOS)**. Mô hình giải thích rất ít sự biến thiên của LOS.

**2. Kết quả kiểm định ANOVA**

* **F-statistic: 2.49**
* **P-value: 0.0217**
* **Kết luận: 👉 Có sự khác biệt có ý nghĩa thống kê giữa các nhóm tuổi.**

**Phân tích sự khác biệt giữa hai kết quả và nhận định chung:**

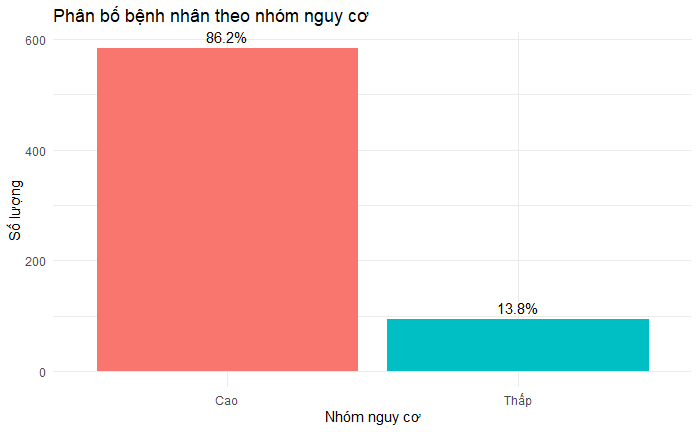
Có một sự khác biệt quan trọng giữa kết quả OLS và ANOVA, và điều này là hợp lý khi xem xét cách thức các phân tích này được thực hiện:

* **ANOVA (ảnh 2) tập trung vào các "nhóm tuổi" (age\_group):**
  + Kết quả ANOVA (p-value = 0.0217 < 0.05) chỉ ra rằng **có sự khác biệt có ý nghĩa thống kê về LOS giữa ít nhất một cặp nhóm tuổi** (ví dụ: nhóm 30-39 có LOS trung bình cao đột biến do outliers như đã phân tích trước đây).
  + ANOVA rất nhạy cảm với sự khác biệt giữa các nhóm, ngay cả khi chỉ có một vài nhóm có giá trị trung bình khác biệt đáng kể.
* **Hồi quy OLS (ảnh 1) sử dụng "tuổi" (age\_years) như một biến liên tục:**
  + Khi sử dụng age\_years như một biến liên tục trong hồi quy, mô hình tìm kiếm một mối quan hệ tuyến tính xuyên suốt tất cả các độ tuổi.
  + Kết quả P>|t| = 0.946 cho thấy không có mối quan hệ tuyến tính đáng kể giữa tuổi và LOS. Điều này phù hợp với biểu đồ tương quan trước đó (hệ số tương quan 0.08) - mối quan hệ là rất yếu.
  + Sự khác biệt được tìm thấy bởi ANOVA có thể là do các trường hợp ngoại lệ (outliers) gây ra sự chênh lệch lớn ở một số nhóm tuổi cụ thể (ví dụ: nhóm 30-39 có LOS cực cao do một vài bệnh nhân), điều mà mô hình hồi quy tuyến tính với biến tuổi liên tục không thể nắm bắt hoặc giải thích tốt.

**Tổng kết chung:**

1. **Về giới tính:** Cả phân tích OLS (p = 0.171) và phân tích ANOVA riêng về giới tính (bạn đã cung cấp trước đó, p = 0.167) đều nhất quán cho thấy **giới tính không phải là yếu tố có ý nghĩa thống kê trong việc giải thích sự khác biệt về thời gian nằm viện.**
2. **Về tuổi:**
   * Khi tuổi được xem xét như một **biến liên tục** trong mô hình hồi quy OLS, nó **không có mối quan hệ tuyến tính có ý nghĩa thống kê** với LOS.
   * Tuy nhiên, khi tuổi được **phân nhóm (age\_group)** và phân tích bằng ANOVA, **có sự khác biệt có ý nghĩa thống kê về LOS giữa ít nhất một vài nhóm tuổi.** Điều này có thể là do sự hiện diện của các trường hợp ngoại lệ (outliers) hoặc các mối quan hệ phi tuyến tính mà mô hình hồi quy tuyến tính đơn giản không thể nắm bắt. Các nhóm tuổi cụ thể có thể có hành vi LOS khác biệt rõ rệt (ví dụ: nhóm 30-39 với các LOS cực đoan).
3. **Hạn chế của mô hình hồi quy OLS:** R-squared cực thấp của mô hình OLS chỉ ra rằng age\_years và Gender\_code không phải là các yếu tố chính giải thích sự biến thiên của LOS. Có rất nhiều yếu tố khác (ví dụ: loại bệnh, mức độ nặng của bệnh, các bệnh kèm theo, loại hình chăm sóc, phương thức ra viện, v.v.) có thể có ảnh hưởng lớn hơn nhiều đến thời gian nằm viện.

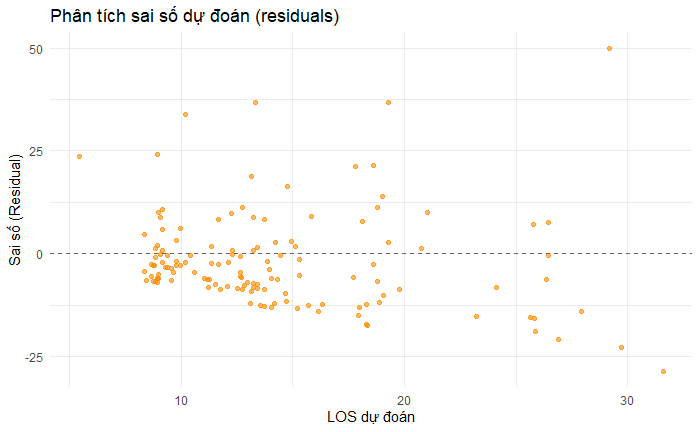
# 3.2. Phân tích theo nhóm nguy cơ cao



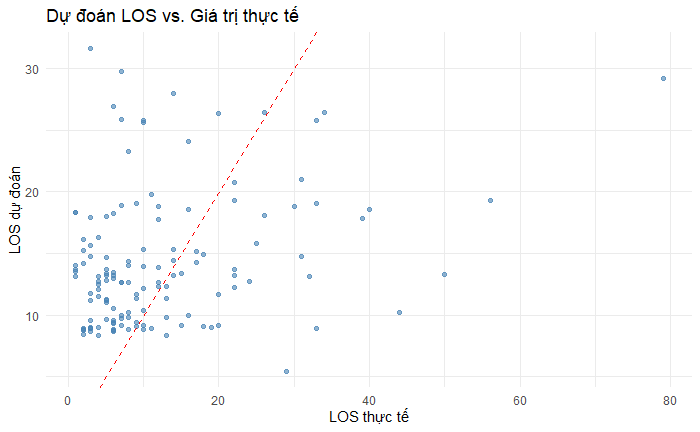
Hình 25: Phân bố bệnh nhân theo nguy cơ

Kết quả cho thấy **giới tính không ảnh hưởng đáng kể đến thời gian nằm viện.** Về tuổi, mặc dù không có mối quan hệ tuyến tính mạnh với LOS, nhưng **có sự khác biệt về thời gian nằm viện giữa các nhóm tuổi cụ thể**, có thể do các trường hợp ngoại lệ hoặc cấu trúc phi tuyến tính của dữ liệu mà mô hình hồi quy tuyến tính không nắm bắt được. Các yếu tố khác ngoài tuổi và giới tính dường như đóng vai trò lớn hơn nhiều trong việc quyết định thời gian nằm viện.

# 3.3. Xây dựng mô hình dự đoán thời gian nằm viện



Hình 26: Phân tích sai số dự đoán (residual)



Hình 27: Dự đoán LOS vs Giá trị thực tế

Nhận xét: Mô hình hồi quy tuyến tính này **dự đoán thời gian nằm viện (LOS) rất kém hiệu quả.** RMSE cao (60.10) và các biểu đồ cho thấy **dự đoán không chính xác và sai số có mẫu hình rõ rệt, vi phạm giả định mô hình.** Cần sử dụng các biến mạnh hơn hoặc loại mô hình khác để dự đoán LOS.

# IV. Hệ thống báo cáo trực tuyến sử dụng Shiny

# 4.1. Mục tiêu xây dựng

Để phục vụ việc phân tích và khai thác dữ liệu bệnh nhân thiếu máu một cách linh hoạt và trực quan, nhóm đã xây dựng **hai phiên bản hệ thống báo cáo trực tuyến** sử dụng ngôn ngữ **R** và thư viện **Shiny**:

* **Phiên bản 1**: sử dụng dữ liệu được nạp sẵn từ file nội bộ trong code.
* **Phiên bản 2**: cho phép người dùng **tải dữ liệu của riêng họ lên**, hệ thống sẽ tự động xử lý và cập nhật các phân tích tương ứng.

Cả hai hệ thống đều có giao diện trực quan, nhiều tab chức năng, biểu đồ động và bảng số liệu tương tác, giúp người dùng không cần viết mã vẫn có thể truy vấn và hiểu dữ liệu.

# 4.2. Cấu trúc và công nghệ sử dụng

* **Ngôn ngữ lập trình**: R
* **Thư viện chính**:  
  shiny, shinydashboard, ggplot2, dplyr, readxl, DT, caret, corrplot, ggpubr
* **Phiên bản 1**:
  + Dữ liệu đọc sẵn từ filtered\_admission.xlsx, diagnosis\_chi\_D500\_D508\_D509.xlsx
  + Phù hợp để minh họa nhanh, không cần thao tác thêm
* **Phiên bản 2**:
  + Cho phép người dùng tải lên **bất kỳ file nhập viện và chẩn đoán phù hợp định dạng**
  + Có thêm nút “Dùng dữ liệu mẫu” để thử nghiệm
  + Tự động lọc bệnh nhân thiếu máu, tính toán nhóm nguy cơ, sinh biểu đồ và mô hình dự đoán
  + Giao diện động, hiển thị các biểu đồ nhân khẩu học, thời gian nằm viện, phân tích nâng cao, kiểm định t-test...

# 4.3. Lý do xây dựng 2 phiên bản

* **Phiên bản 1** giúp trình bày nhanh cho người không có dữ liệu.
* **Phiên bản 2** là mô hình linh hoạt hơn, có thể áp dụng trong thực tế với **bất kỳ bộ dữ liệu ICD phù hợp nào**, không bị cố định.
* Việc có khả năng cho phép **người dùng tự tải dữ liệu lên** là bước tiến gần hơn tới một công cụ dùng trong môi trường y tế thật, giúp cán bộ y tế có thể áp dụng với dữ liệu thực tế từ các bệnh viện khác nhau.

# 4.4. Lợi ích của hệ thống

* Hỗ trợ ra quyết định dựa trên dữ liệu thực tế;
* Cho phép cán bộ y tế dễ dàng phân tích bệnh thiếu máu mà không cần kỹ năng lập trình;
* Có thể tái sử dụng, mở rộng với các bệnh khác bằng cách thay đổi bộ ICD tương ứng;
* Giao diện đơn giản, chạy trên trình duyệt, dễ triển khai nội bộ hoặc chia sẻ trong nhóm nghiên cứu.

# 4.5. So sánh hai phiên bản hệ thống báo cáo Shiny

Nhóm đã phát triển hai phiên bản ứng dụng Shiny nhằm đáp ứng các nhu cầu khác nhau của người sử dụng. Bảng dưới đây tóm tắt sự khác biệt về chức năng, mục tiêu và mức độ linh hoạt giữa hai phiên bản:

| **Tiêu chí** | **Phiên bản 1: Cố định dữ liệu** | **Phiên bản 2: Cho phép tải dữ liệu** |
| --- | --- | --- |
| **Mục đích chính** | Minh họa nhanh, chạy demo với dữ liệu được lập trình sẵn | Phân tích linh hoạt với bất kỳ bộ dữ liệu phù hợp mà người dùng tải lên |
| **Dữ liệu** | Đọc từ các file .xlsx được nạp sẵn trong code | Người dùng tải lên thủ công qua giao diện |
| **Tính tương tác với người dùng** | Giới hạn, chỉ xem được dữ liệu mẫu | Cao – người dùng tự chọn dữ liệu để phân tích |
| **Tính tái sử dụng** | Chỉ áp dụng cho bộ dữ liệu cố định | Có thể áp dụng với nhiều bệnh viện và tập dữ liệu ICD khác nhau |
| **Tính năng đặc biệt** | Không có nút “Tải lên” hay chọn file | Có chức năng fileInput, nút “Dùng dữ liệu mẫu” |
| **Tự động xử lý dữ liệu** | Có xử lý nhưng cố định ID bệnh nhân | Tự động lọc bệnh nhân theo ICD10: D500, D508, D509 sau khi tải dữ liệu |
| **Phù hợp cho đối tượng** | Người xem nhanh báo cáo mẫu, không cần thao tác | Cán bộ y tế hoặc nhà nghiên cứu muốn áp dụng cho dữ liệu thực tế |
| **Rủi ro lỗi từ phía người dùng** | Thấp – dữ liệu luôn đúng định dạng | Có – nếu tải sai định dạng hoặc thiếu cột bắt buộc |

# 4.6. Kết luận về lựa chọn triển khai

Trong thực tế, phiên bản 2 là **linh hoạt và thực tiễn hơn**, phù hợp với yêu cầu của một hệ thống báo cáo y tế triển khai rộng rãi tại bệnh viện. Tuy nhiên, phiên bản 1 vẫn rất hữu ích cho việc trình diễn, giảng dạy hoặc kiểm thử tính năng một cách nhanh chóng.

Nhóm khuyến nghị duy trì **cả hai phiên bản**:

* **Phiên bản 1** dùng để đào tạo nội bộ, chạy demo.
* **Phiên bản 2** dùng cho thực hành phân tích dữ liệu thật tại các cơ sở y tế.

# V. Đánh giá chất lượng bộ dữ liệu

* **Thiếu giá trị (missing)**: Có một số dòng bị thiếu thông tin về tuổi hoặc giới tính (~2.1%).
* **Ngoại lệ (outliers)**: Một số ca có thời gian nằm viện > 500 ngày, cần được kiểm tra kỹ (có thể do lỗi nhập liệu).
* **Chưa đồng nhất mã ICD/DRG**: Một số mã DRG hoặc mã ICD không khớp hoàn toàn với mô tả, cần chuẩn hóa thêm.
* **Sự vắng mặt của mã nhập viện M (thông thường)**: Có thể là lỗi lọc dữ liệu hoặc thiếu biến trong bộ dữ liệu gốc.

Tổng thể, chất lượng dữ liệu **chấp nhận được cho nghiên cứu mô tả**, nhưng nếu dùng cho dự đoán/phân tích chính sách, cần cải thiện độ đầy đủ và tính chuẩn hóa

# VI. Lời khuyên dành cho người sử dụng dữ liệu thiếu máu do thiếu sắt

Khi sử dụng dữ liệu về thiếu máu do thiếu sắt, cần đặc biệt lưu ý **hạn chế mã hóa ICD-10 dưới dạng không xác định** như D50.9 trong trường hợp đã có đầy đủ thông tin về nguyên nhân bệnh. Việc sử dụng đúng mã sẽ **nâng cao độ chính xác và khả năng khai thác dữ liệu**, đặc biệt trong các phân tích nguyên nhân hoặc xu hướng dịch tễ.

Trong quá trình xây dựng kế hoạch truyền thông và can thiệp y tế, việc **phân tích dữ liệu theo nhóm tuổi và giới tính** đóng vai trò then chốt trong việc xác định các nhóm nguy cơ cao. Điều này giúp thiết kế các chương trình giáo dục sức khỏe và can thiệp một cách **phù hợp, có mục tiêu và hiệu quả**.

Ngoài ra, việc **tích hợp dữ liệu bệnh án với kết quả khám sức khỏe học đường** sẽ hỗ trợ theo dõi xu hướng sức khỏe dài hạn, phát hiện sớm các biểu hiện thiếu máu và **đánh giá tác động của các biện pháp can thiệp**, từ đó nâng cao chất lượng chăm sóc sức khỏe cộng đồng trong cả môi trường trường học lẫn xã hội.

# THẢO LUẬN BÀI BÁO: DXPR – CÔNG CỤ CHUẨN HOÁ DỮ LIỆU ICD trong nghiên cứu HER

**Tài liệu tham khảo**

Tseng YJ, Chiu HJ, Chen CJ. [*dxpr: an R package for generating analysis-ready data from electronic health records—diagnoses and procedures*](https://pubmed.ncbi.nlm.nih.gov/34141876/). PeerJ Comput Sci. 2021;7:e520. doi:10.7717/peerj-cs.520.

# I. Giới thiệu

Gói R **dxpr**, được công bố vào năm 2021 bởi Tseng, C. H., và cộng sự, là một công cụ đột phá được phát triển với mục tiêu chính là **"đơn giản hóa và tăng tốc quá trình phân tích dữ liệu hồ sơ sức khỏe điện tử (EHR) và dữ liệu khiếu nại y tế, đặc biệt trong bối cảnh nghiên cứu đa trung tâm."** Sự ra đời của **dxpr** nhằm giải quyết một trong những thách thức lớn nhất trong nghiên cứu y học hiện đại: việc chuẩn bị dữ liệu lâm sàng từ các nguồn EHR thô, vốn thường rất phức tạp và đòi hỏi nhiều công sức tiền xử lý thủ công.

**dxpr** tập trung vào ba bước quan trọng trong chu trình xử lý dữ liệu: **tích hợp dữ liệu**, **làm sạch và chuẩn bị dữ liệu (data wrangling)**, và **trực quan hóa dữ liệu chẩn đoán và thủ thuật**. Bằng cách tự động hóa nhiều tác vụ lặp đi lặp lại và phức tạp, gói này giúp các nhà nghiên cứu lâm sàng nhanh chóng có được bộ dữ liệu "sẵn sàng để phân tích" (analysis-ready dataset) mà không cần phải viết hàng trăm dòng mã tiền xử lý tùy chỉnh. Điều này đặc biệt hữu ích khi làm việc với các bộ dữ liệu lớn, đa dạng về định dạng và nguồn gốc, đồng thời thúc đẩy khả năng hợp tác nghiên cứu giữa các trung tâm y tế khác nhau.

# II. Chức năng chính của gói R dxpr

**dxpr** cung cấp một bộ các hàm chức năng cốt lõi được thiết kế để xử lý và chuẩn hóa các mã chẩn đoán và thủ thuật, đáp ứng các yêu cầu khắt khe của phân tích dịch tễ học và lâm sàng.

## 2.1 Chuẩn hóa mã ICD

Tính nhất quán của mã ICD (International Classification of Diseases) là cực kỳ quan trọng cho mọi phân tích dữ liệu y tế. **dxpr** giải quyết vấn đề này bằng cách cung cấp các công cụ linh hoạt để chuyển đổi định dạng và kiểm tra lỗi:

* **Chuyển đổi định dạng mã ICD**:
  + **icdDxShortToDecimal()**: Chức năng này cho phép chuyển đổi mã ICD từ định dạng "short" (ví dụ: I10) sang định dạng "decimal" (ví dụ: I10.0). Định dạng decimal thường được yêu cầu bởi một số công cụ phân loại hoặc cơ sở dữ liệu.
  + **icdDxDecimalToShort()**: Ngược lại, chức năng này chuyển đổi mã ICD từ định dạng decimal sang short. Khả năng chuyển đổi qua lại này đặc biệt hữu ích để đáp ứng các tiêu chí cụ thể của các phương pháp phân loại bệnh như **PheWAS (Phenome-Wide Association Study)**, **CCS (Clinical Classification Software)**, hoặc khi tính toán các chỉ số bệnh đồng mắc như **Charlson/Elixhauser**.
* **Kiểm tra và cảnh báo lỗi tự động**: Một điểm mạnh khác của **dxpr** là khả năng tự động kiểm tra và cảnh báo về các lỗi mã hóa. Gói sẽ nhận diện các mã có định dạng sai (ví dụ: ký tự không hợp lệ, độ dài không đúng) hoặc sử dụng phiên bản ICD không chính xác (ví dụ: lẫn lộn ICD-9 và ICD-10 trong cùng một cột dữ liệu). Hệ thống cảnh báo này giúp người dùng nhanh chóng xác định và khắc phục các vấn đề về chất lượng dữ liệu, tiết kiệm đáng kể thời gian và công sức so với việc kiểm tra thủ công.

## 2.2. Nhóm mã ICD

Để biến các mã ICD chi tiết thành các nhóm bệnh hoặc tình trạng có ý nghĩa lâm sàng, **dxpr** cung cấp nhiều phương pháp nhóm mã khác nhau, phù hợp với các mục tiêu nghiên cứu đa dạng:

* **Nhóm mã chẩn đoán**:
  + **CCS cấp đơn (icdDxToCCS)**: Đây là phương pháp nhóm các mã ICD rất chi tiết thành các danh mục bệnh lý lâm sàng rộng hơn. Ví dụ, nhiều mã ICD khác nhau cho các loại bệnh tim mạch có thể được nhóm lại thành một nhóm CCS chung là "Bệnh tim mạch".
  + **CCS đa cấp (icdDxToCCSLvl)**: Cung cấp khả năng nhóm mã ICD theo nhiều cấp độ chi tiết, từ tổng quát đến cụ thể. Điều này cho phép nhà nghiên cứu linh hoạt lựa chọn mức độ phân loại phù hợp với câu hỏi nghiên cứu của mình.
  + **PheWAS (icdDxToPheWAS)**: Chuyển đổi các mã ICD thành các "phecode" (mã kiểu hình bệnh lý), là các nhóm bệnh được thiết kế đặc biệt cho các nghiên cứu PheWAS nhằm khám phá mối liên hệ giữa các biến thể di truyền và các đặc điểm bệnh lý.
  + **Comorbidity Charlson/Elixhauser (icdDxToComorbid)**: Tính toán các chỉ số bệnh tật đồng mắc phổ biến như Charlson Comorbidity Index và Elixhauser Comorbidity Index. Các chỉ số này rất quan trọng trong việc điều chỉnh nguy cơ và đánh giá kết cục ở bệnh nhân, đặc biệt trong nghiên cứu dịch tễ học và kết cục sức khỏe.
* **Nhóm mã thủ thuật**: Tương tự như mã chẩn đoán, **dxpr** cũng hỗ trợ nhóm các mã thủ thuật (ICD-9 PCS hoặc ICD-10 PCS) thông qua các hàm như icdPrToCCS, icdPrToCCSLvl, và icdPrToProcedureClass. Điều này cho phép phân loại và phân tích các can thiệp y tế một cách có hệ thống.
* **Định dạng kết quả đầu ra**: Các chức năng nhóm mã của **dxpr** trả về hai loại bảng kết quả:
  + Một bảng chi tiết theo từng bản ghi (record-level), giữ lại thông tin cụ thể về từng lần chẩn đoán hoặc thủ thuật.
  + Một bảng tóm tắt, cung cấp thông tin tổng hợp hữu ích như ngày chẩn đoán/thủ thuật đầu tiên và cuối cùng, tổng số lần chẩn đoán/thủ thuật, và khoảng thời gian giữa các sự kiện. Bảng tóm tắt này đặc biệt hữu ích cho các phân tích cấp bệnh nhân.

# III. Xử lý dữ liệu lâm sàng (Data Wrangling)

Quá trình "data wrangling" (làm sạch và chuẩn bị dữ liệu) là bước tiêu tốn nhiều thời gian nhất trong phân tích EHR. **dxpr** cung cấp một tập hợp các công cụ mạnh mẽ để tự động hóa các tác vụ này, giúp dữ liệu trở nên "sạch" và sẵn sàng cho phân tích.

* **selectCases()**: Chức năng này cho phép lọc và chọn các nhóm bệnh nhân cụ thể (nhóm ca và nhóm đối chứng) dựa trên các tiêu chí phức tạp. Các tiêu chí có thể bao gồm sự hiện diện của một mã chẩn đoán hoặc thủ thuật cụ thể, số lần xuất hiện của mã đó, và khoảng thời gian mà các sự kiện này xảy ra. Đây là công cụ thiết yếu để xây dựng các đoàn hệ nghiên cứu hoặc so sánh giữa các nhóm.
* **getEligiblePeriod()**: Xác định khoảng thời gian mà dữ liệu của mỗi bệnh nhân được coi là hợp lệ và đầy đủ. Hàm này tính toán ngày ghi nhận dữ liệu đầu tiên và cuối cùng cho mỗi bệnh nhân, giúp loại bỏ những bệnh nhân không đủ thời gian theo dõi hoặc có dữ liệu không liên tục, đảm bảo tính hợp lệ cho các phân tích dài hạn.
* **splitDataByDate()**: Cho phép phân tách dữ liệu của mỗi bệnh nhân xung quanh một "ngày chỉ định" (index date) – ví dụ, ngày chẩn đoán bệnh ung thư, ngày bắt đầu điều trị. Chức năng này có thể tạo ra các "cửa sổ" thời gian tùy chọn (ví dụ: 30 ngày trước và 90 ngày sau ngày chỉ định) để nghiên cứu các sự kiện xảy ra trong các giai đoạn cụ thể của bệnh.
* **getConditionEra()**: Một trong những chức năng đột phá của **dxpr**. Nó gom các lần chẩn đoán gần nhau của cùng một bệnh hoặc tình trạng thành một "condition era" (thời kỳ bệnh tật). Ví dụ, nếu một bệnh nhân đến khám nhiều lần trong một thời gian ngắn và đều được chẩn đoán cùng một bệnh, thay vì coi đó là nhiều lần chẩn đoán riêng lẻ, getConditionEra() sẽ tổng hợp chúng thành một chuỗi lâm sàng liên tục. Điều này giúp phản ánh chính xác hơn quá trình bệnh lý của bệnh nhân và tạo ra dữ liệu phù hợp hơn cho các mô hình phân tích chuỗi thời gian.
* **groupedDataLongToWide()**: Chuyển đổi cấu trúc dữ liệu từ dạng "dài" (long format), nơi mỗi hàng là một sự kiện, sang dạng "rộng" (wide format), nơi mỗi hàng là một bệnh nhân và các biến được trải rộng ra thành các cột riêng biệt. Định dạng rộng thường được ưa chuộng cho các phân tích thống kê truyền thống và dễ dàng kết hợp với các gói R khác như tableone để tạo các bảng đặc điểm cơ bản của quần thể nghiên cứu.

# IV. Xử lý dữ liệu thủ thuật (Procedure Data)

Gói **dxpr** không chỉ giới hạn ở dữ liệu chẩn đoán mà còn mở rộng khả năng mạnh mẽ của mình để xử lý dữ liệu thủ thuật, vốn cũng rất quan trọng trong việc mô tả quá trình điều trị và can thiệp y tế.

* **Tương tự như dữ liệu chẩn đoán**: **dxpr** cung cấp các chức năng tương tự cho dữ liệu thủ thuật, bao gồm khả năng **chuẩn hóa mã thủ thuật** (ví dụ: ICD-9 PCS hoặc ICD-10 PCS) và **hỗ trợ nhóm thủ thuật** theo các hệ thống phân loại như CCS hoặc các lớp thủ thuật (procedure class) tùy chỉnh. Điều này đảm bảo tính nhất quán và khả năng phân tích dữ liệu thủ thuật một cách có hệ thống.
* **Phân tích chuỗi thủ thuật**: Tương tự như getConditionEra(), **dxpr** cũng cho phép phân tích chuỗi các thủ thuật liên tiếp. Điều này rất hữu ích trong việc xây dựng các "chuỗi thủ thuật" hoặc "thời kỳ can thiệp", giúp nhà nghiên cứu hiểu rõ hơn về lộ trình điều trị, hiệu quả của các can thiệp liên tiếp, hoặc đánh giá các biến chứng liên quan đến một chuỗi thủ thuật cụ thể.

# V. Trực quan hóa và kiểm định

Để hỗ trợ quá trình kiểm tra chất lượng dữ liệu, khám phá dữ liệu ban đầu và trình bày kết quả, **dxpr** cung cấp các công cụ trực quan hóa đơn giản nhưng hiệu quả.

* **plotICDError()**: Hàm này tạo ra một **biểu đồ Pareto** hiển thị các mã ICD lỗi phổ biến nhất trong tập dữ liệu của bạn. Biểu đồ Pareto là một công cụ mạnh mẽ trong kiểm soát chất lượng, giúp nhà nghiên cứu nhanh chóng xác định "nguyên nhân gốc rễ" của các vấn đề dữ liệu, ưu tiên những lỗi cần được xử lý trước khi tiến hành phân tích chuyên sâu.
* **plotDiagCat()**: Chức năng này tạo ra các **biểu đồ thanh** giúp so sánh tần suất hoặc tỷ lệ của các nhóm chẩn đoán (ví dụ: các nhóm CCS hoặc PheWAS) giữa các nhóm nghiên cứu khác nhau (ví dụ: nhóm ca bệnh và nhóm đối chứng). Quan trọng hơn, plotDiagCat() còn tích hợp khả năng thực hiện **kiểm định thống kê** (ví dụ: kiểm định Chi-square hoặc kiểm định Fisher's exact) trực tiếp trên biểu đồ, cung cấp giá trị p để đánh giá sự khác biệt có ý nghĩa thống kê giữa các nhóm. Ví dụ như trong bài báo gốc, hàm này có thể được dùng để so sánh tỷ lệ mắc các bệnh liên quan đến PDA (Pulmonary Disease Associations) giữa các nhóm.

# VI. Phân tích hiệu suất (Performance)

Hiệu suất là một yếu tố then chốt đối với mọi gói xử lý dữ liệu lớn. Bài báo của Tseng et al. đã trình bày các thử nghiệm chi tiết để đánh giá khả năng xử lý của **dxpr** trên một bộ dữ liệu quy mô lớn, chứng minh hiệu quả của nó trong môi trường thực tế.

* **Quy mô dữ liệu thử nghiệm**: Thử nghiệm được thực hiện trên một bộ dữ liệu ấn tượng bao gồm **953.294 bệnh nhân** và tổng cộng **7.948.418 bản ghi** về chẩn đoán và thủ thuật. Đây là một quy mô dữ liệu điển hình cho các nghiên cứu EHR thực tế.
* **Cấu hình phần cứng**: Các thử nghiệm được thực hiện trên một máy tính có cấu hình tiêu chuẩn cao cấp: **CPU Core i7 và RAM 64 GB**. Điều này cho thấy **dxpr** có thể hoạt động hiệu quả trên các hệ thống máy tính phổ biến trong các trung tâm nghiên cứu.
* **Thời gian xử lý trung bình các chức năng cốt lõi**:
  + selectCases(): Khoảng **238 giây** (~4 phút).
  + getEligiblePeriod(): Khoảng **1.1 giây**.
  + splitDataByDate(): Khoảng **6.5 giây**.
  + getConditionEra(): Khoảng **372 giây** (~6 phút).
  + Chuyển đổi dạng dữ liệu (groupedDataLongToWide()): Khoảng **3.8 giây**.

Những con số này chứng minh rằng **dxpr** có khả năng xử lý các tập dữ liệu EHR lớn một cách **khá hiệu quả và nhanh chóng**. Hiệu suất mạnh mẽ này làm cho **dxpr** trở thành một công cụ lý tưởng để hỗ trợ các phân tích dữ liệu quy mô lớn, bao gồm cả các nghiên cứu quốc tế hoặc hợp tác liên cơ sở, nơi việc xử lý dữ liệu tập trung thường là một nút thắt cổ chai.

# VII. Kết luận và hướng phát triển

**dxpr** đại diện cho một bước tiến quan trọng trong việc đơn giản hóa và tăng tốc phân tích dữ liệu EHR. Mặc dù có những hạn chế nhất định, tiềm năng phát triển của gói này là rất lớn.

## 7.1. Ưu điểm:

* **Tiết kiệm thời gian xử lý đáng kể**: Bằng cách tự động hóa nhiều tác vụ làm sạch, chuẩn hóa và nhóm dữ liệu, **dxpr** giúp các nhà nghiên cứu giảm thiểu thời gian dành cho tiền xử lý, cho phép họ tập trung hơn vào việc phân tích và diễn giải kết quả.
* **Tạo điều kiện cho hợp tác nghiên cứu liên viện**: Khả năng chuẩn hóa mã ICD và cấu trúc dữ liệu theo các tiêu chuẩn chung giúp dễ dàng chia sẻ và tích hợp dữ liệu từ nhiều nguồn khác nhau, thúc đẩy các nghiên cứu đa trung tâm và hợp tác quốc tế.
* **Script cấu trúc rõ ràng và dễ chia sẻ**: Gói **dxpr** được thiết kế với các hàm có tên rõ ràng và luồng công việc logic, làm cho các đoạn mã trở nên dễ hiểu, dễ tái sử dụng và dễ dàng chia sẻ giữa các thành viên trong nhóm nghiên cứu.

**Hạn chế:**

* **Phụ thuộc vào độ chính xác của mã ICD tại nguồn**: Mặc dù **dxpr** có cơ chế kiểm tra lỗi, nhưng chất lượng đầu ra cuối cùng vẫn phụ thuộc nhiều vào chất lượng và tính chính xác của dữ liệu gốc tại các hệ thống EHR. Nếu dữ liệu nguồn chứa nhiều lỗi mã hóa cơ bản, việc làm sạch sẽ vẫn là một thách thức.
* **Cần cân nhắc khi chuyển đổi giữa ICD-9 và ICD-10 trong cùng tập dữ liệu**: Việc xử lý và tích hợp dữ liệu có chứa cả ICD-9 và ICD-10 đòi hỏi sự cẩn trọng đặc biệt để tránh sai sót và đảm bảo tính nhất quán của phân loại bệnh tật qua các phiên bản.
* **Không xử lý đa dạng dữ liệu**: Hiện tại, **dxpr** tập trung chủ yếu vào dữ liệu chẩn đoán và thủ thuật. Nó không có chức năng tích hợp để xử lý các loại dữ liệu EHR khác như thông tin về thuốc men, kết quả xét nghiệm, chỉ số sinh tồn, hoặc ghi chú y tế. Tuy nhiên, hạn chế này có thể được khắc phục bằng cách **kết hợp dxpr với các gói R chuyên biệt khác** như AdhereR (cho dữ liệu tuân thủ thuốc) hoặc các gói chuyên về dữ liệu lab/kết quả xét nghiệm.

## 7.2. Hướng phát triển:

Nhóm phát triển đã vạch ra những hướng đi đầy hứa hẹn cho tương lai của **dxpr**:

* **Mở rộng tính năng xử lý**: Phát triển các mô-đun mới để xử lý các loại dữ liệu EHR đa dạng hơn, bao gồm thông tin về thuốc men, liều lượng, lịch sử sử dụng thuốc, và phân tích ghi chú y tế (ví dụ: bằng các kỹ thuật xử lý ngôn ngữ tự nhiên - NLP).
* **Tích hợp giao diện người dùng (Shiny)**: Xây dựng một giao diện người dùng đồ họa (GUI) dựa trên gói Shiny của R. Điều này sẽ cho phép người dùng thực hiện các phân tích phức tạp mà không cần viết mã R, tạo ra các báo cáo tự động và báo cáo động (dynamic reports) về kết quả phân tích. Giao diện này sẽ làm cho **dxpr** trở nên thân thiện và dễ tiếp cận hơn với các nhà nghiên cứu không có kinh nghiệm lập trình sâu rộng.

# VII. Tổng kết

Gói **dxpr** là một giải pháp toàn diện và hiệu quả cho các nghiên cứu sử dụng dữ liệu EHR, đặc biệt là trong việc xử lý và chuẩn bị dữ liệu chẩn đoán và thủ thuật. Nó đáp ứng nhu cầu cấp thiết về việc tạo ra các tập dữ liệu "analysis-ready" một cách nhanh chóng, dễ dàng chia sẻ và chuẩn hóa giữa các nguồn dữ liệu khác nhau. Với khả năng tự động hóa các tác vụ tiền xử lý phức tạp, **dxpr** không chỉ tiết kiệm thời gian và công sức mà còn nâng cao chất lượng và độ tin cậy của các phân tích lâm sàng. Đây là một công cụ rất phù hợp cho bất kỳ dự án nào muốn tối ưu hóa quy trình làm việc với dữ liệu y tế lớn.