Những sai sót và thiếu sót phổ biến trong phân tích dữ liệu

(trích từ sách "**Phân tích dữ liệu với R: Hỏi và Đáp**", Nhà xuất bản Tổng Hợp 2017)

Phân tích thống kê là một phần không thể thiếu được trong các nghiên cứu y khoa, nhất là nghiên cứu lâm sàng và dịch tễ học. Thống kê đã được ứng dụng trong y học từ những năm trong thập niên 1930s, nhưng thật ra từ thế kỉ 19 người ta cũng đã nghĩ đến việc sử dụng các thuật phân tích thống kê trong thử nghiệm lâm sàng. Mặc dù đã trải qua hơn 1 thế kỉ ứng dụng, nhưng cho đến ngày nay vẫn còn rất nhiều sai sót về phân tích thống kê trong các công trình nghiên cứu y học. Một số sai sót không ảnh hưởng gì đến kết luận của nghiên cứu, nhưng nhiều sai sót mang tính hệ thống thì có khi làm cho công trình nghiên cứu có một ý nghĩa hoàn toàn khác với kết luận của tác giả.

Để khắc phục tình trạng sai sót về phân tích thống kê trong nghiên cứu y học, nhiều nhóm trên thế giới đã xuất bản những "phác đồ", những hướng dẫn trong cách phân tích và trình bày kết quả phân tích dữ liệu. Đây là một nỗ lực trong thực hành y học thực chứng, bởi vì y học thực chứng dựa vào những công trình nghiên cứu có chất lượng và chứng cứ phải chính xác. Trong bối cảnh nghiên cứu y học ở Việt Nam, y học thực chứng vẫn còn trong giai đoạn đầu, và nhìn qua những bài báo khoa học rất dễ nhận ra nhiều sai sót về phân tích dữ liệu trong mỗi bài báo. Đó cũng là một trong những nguyên nhân dẫn đến chất lượng nghiên cứu y học ở Việt Nam không có phẩm chất cao. Chúng ta cần phải khắc phục tình trạng yếu kém này.

Phân tích thống kê có thể chia thành hai lĩnh vực chính: phân tích mô tả và phân tích suy luận. Phân tích mô tả quan tâm đến cách mô tả dữ liệu từ một mẫu hoặc từ một công trình nghiên cứu. Phân tích suy luận bao gồm các phương pháp phân tích cách ước tính, suy luận, kiểm định giả thuyết khoa học. Trong chương này, tôi sẽ trình bày những sai sót phổ biến nhất về phân tích mô tả và phân tích suy luận trong các nghiên cứu y học trên thế giới và Việt Nam, với hi vọng rằng những sai sót hoặc thiếu sót (sẽ gọi chung là "thiếu sót") này sẽ giảm đi trong tương lai, và chất lượng nghiên cứu khoa học sẽ được nâng cao.

Thiếu sót 1: Không định nghĩa biến phân tích rõ ràng.

Đặc tính của khoa học là cân, đo, đong, đếm. Nhà nghiên cứu cần phải cho người đọc (và công chúng) biết những biến số họ đo lường là gì, và phải cung cấp định nghĩa của những biến đó sao cho người đọc có thể hiểu được. Chẳng hạn như khi nói đến mật độ xương (bone mineral density – BMD), nhà nghiên cứu phải định nghĩa BMD là gì, đo ở vị trí nào trong cơ thể, đơn vị đo lường là gì, và đo bằng phương pháp hay phương tiện gì. Hay như huyết áp, nhà nghiên cứu phải cung cấp định nghĩa ngưỡng giá trị nào là "cao huyết áp" và ngưỡng nào là "bình thường" và dựa trên cách phân

loại nào (điều này đặc biệt quan trọng trong những trường hợp cách phân loại còn chưa thống nhất ở thời điểm hiện tại). Tương tự, khi đề cập đến béo phì (obesity), nhà nghiên cứu phải định nghĩa thế nào là béo phì, và dùng chỉ số nào để định nghĩa. Chẳng hạn như BMI trên 27.5 (ở người Á châu) hay trên 30 (ở người Âu châu) được xem là béo phì.

Đối với các biến liên quan đến khái niệm hoặc hành vi (behavior) vấn đề định nghĩa có thể khó hơn vì khó đo lường. Chẳng hạn như trầm cảm được định nghĩa bằng thang điểm Zung Depression Inventory (ZDI) trên 50, nhưng biến số này phản ảnh trầm cảm chính xác độ nào thì là một vấn đề còn trong vòng tranh cãi. Trong một cuộc điều tra qui mô ở Mĩ, một cá nhân được xem là "đang hút thuốc lá" là người hút thuốc lá trong vòng 30 ngày trước khi tham gia cuộc điều tra. Mặc dù định nghĩa này không hiển nhiên như chúng ta mong muốn, nhưng đó là một định nghĩa mang tính "operational", và nhà nghiên cứu phải phát biểu, cho dù chúng ta có thể không đồng ý với định nghĩa đó.

Thiếu sót 2: Không cung cấp độ đo lường cho từng biến số.

Độ đo lường (level of measurement) là một thông tin quan trọng cho phân tích thống kê. Trong lí thuyết đo lường, người ta phân biệt ba loại biến số: định danh (nominal), định cấp độ (ordinal), và liên tục (continuous). Ở mức độ thấp nhất là những dữ liệu mang tính định danh, tức những biến bao gồm hai hay hơn hai loại (nam hay nữ), hoặc tên (đạo Phật, đạo Công giáo), phân loại nhưng không có tính thứ tự cao thấp (như nghề nghiệp). Loại máu A, B, AB, hoặc O cũng được xem là dữ liệu định danh.

Các dữ liệu định cấp bao gồm thể loại có thứ tự cao thấp và có thể xếp hạng. chẳng hạn như một cá nhân có thể là thấp, trung bình, hay cao. Chúng ta có thể không biết chính xác chiều cao của bệnh nhân, nhưng chúng ta biết bệnh nhân đó thuộc nhóm cao, trung bình hay thấp. Các biến liên tục có giá trị chính xác hơn các biến định cấp và định danh. Những biến số như chiều cao (đo bằng cm), cân nặng (kg), huyết áp (mmHg), mật độ xương (g/cm^2), v.v. được xem là những dữ liệu liên tục. Dữ liệu liên tục là những dữ liệu có độ chính xác cao nhất trong 3 nhóm đo lường.

Nhà nghiên cứu cần phải nói rõ biến phân tích thuộc loại nào trong bài báo khoa học. Chẳng hạn như huyết áp của một bệnh nhân có thể chia thành hai nhóm (tăng hay không tăng), hoặc như là một biến phân cấp (huyết áp thấp (hypotensive), bình thường (normotensive), và tăng (hypertensive)), hoặc như là một biến liên tục (đơn vị là mmHg). Đây là vấn đề quan trọng, bởi vì đặc tính của biến số quyết định phương pháp phân tích. Do đó, nếu nhà nghiên cứu không định nghĩa và không mô tả biến phân tích rõ ràng, thì người đọc sẽ không lĩnh hội được kết quả nghiên cứu có ý nghĩa gì.

Thiếu sót 3: "Cắt" biến liên tục thành nhiều nhóm một cách tùy tiện.

Thỉnh thoảng, để đơn giản hóa các phân tích thống kê, nhà nghiên cứu có thể chia các biến liên tục thành nhiều nhóm. Chẳng hạn như tỉ trọng cơ thể (body mass index hay BMI) có thể chia thành 4 nhóm: béo phì, quá cân, bình thường, và thiếu cân. Đó là cách phân nhóm có lí do khoa học và đã được minh chứng theo thời gian là khá hợp lí (dù còn nhiều tranh cãi). Do đó, một phân tích dựa trên cách phân nhóm có cơ sở khoa học như thế ít khi nào bị chất vấn.

Nhưng có nhiều trường hợp nhà nghiên cứu chia nhóm một cách tùy tiện, hoàn toàn không theo một qui ước hay dựa trên một cơ sở khoa học nào cả. Chẳng hạn như có người chia độ tuổi thành nhiều nhóm theo 5 tuổi (0-4, 5-9, 10-14, v.v.), lại có khi chia thành nhóm theo 10 tuổi (0-9, 10-19, 20-29, v.v.). Tương tự, người ta cắt những biến số sinh hoá, thậm chí biến số kinh tế (như thu nhập) ra thành những nhóm độc lập nhau. Chẳng hạn như nếu chúng ta chia tuổi thành 4 nhóm như 0-9, 10-19, 20-29, tức là chúng ta sẽ có 3 nhóm. Nhưng thử tượng một người có tuổi 2 và người tuổi 18, cách nhau đến 12 năm tuổi, nhưng khi phân tích trên nhóm thì chỉ có 1 và 2! Chúng ta mất thông tin. Do đó, cách phân nhóm như thế thường dẫn đến những sai sót trong phân tích, mà lại bị phê bình là thiếu khoa học tính.

Phân chia một biến liên tục thành một biến không liên tục qua phân nhóm như vừa đề cập là một cách làm … phi khoa học. So với các biến định danh (nominal variable), thì biến liên tục (continuous variable) là biến có giá trị chính xác cao nhất. Vì độ chính xác cao, nên các biến liên tục cung cấp nhiều thông tin nhất. Ngược lại, khi cắt biến liên tục thành từng nhóm nhỏ cũng có nghĩa là giảm lượng thông tin của biến số, vì giảm độ chính xác của biến số. Khi biến số bị giảm lượng thông tin, thì kết quả phân tích trên các biến như thế cũng có thể dẫn đến sai lầm.

Khi được hỏi tại sao chia thành từng nhóm nhỏ như thế, thì câu trả lời là để tính mức độ ảnh hưởng. Mức độ ảnh hưởng (effect size) ở đây là tỉ số odds, tỉ số nguy cơ (relative risk – RR). Tại sao nghĩ như thế? Tại vì họ nghĩ rằng không thể tính RR cho các biến liên tục. Nhưng rất tiếc rằng đó là một hiểu sai. Chúng ta vẫn có thể tính RR cho biến liên tục. Chẳng hạn như thay vì tính cho mỗi năm, chúng ta có thể tính cho mỗi 5 năm. Những ai học R có thể biết hàm sau đây

```
glm(death ~ age, family=binomial, data=dd)
```

là tính tỉ số odds tử vong cho mỗi năm tuổi. Nhưng nếu chúng ta muốn tính cho mỗi 5 năm tuổi, thì hàm sẽ là:

```
glm(death ~ I(age/5), family=binomial, data=dd)
```

Nói tóm lại, không nên cắt một biến số liên tục ra thành nhiều nhóm nhỏ. Đó có thể xem là một hình thức "tra tấn dữ liệu" (data torture), một 'bad practice' trong nghiên cứu khoa học. Nếu biến số liên tục có phân bố chuẩn hay gần chuẩn thì nên sử dụng giá trị gốc. Tuy nhiên, việc chia nhóm có thể chấp nhận được nếu phân bố có dạng multimodal (tức có nhiều đỉnh) hoặc cách phân nhóm có ý nghĩa thực tế.

Thiếu sót 4: Không xác định rằng số liệu trong phân tích phương sai (ANOVA) và kiểm định t đáp ứng các giả định thống kê.

Phân tích phương sai, hay một phiên bản đơn giản hơn là kiểm định t, dựa vào giả định rằng số liệu phải tuân theo luật phân phối chuẩn (normal distribution), độc lập với nhau (independence), và phương sai giữa các nhóm không khác nhau (homogeneity). Nhưng trong thực tế sinh học, nhiều biến số không đáp ứng những giả định trên. Nếu nhà nghiên cứu bất chấp các giả định và phân tích thì kết quả có thể không đúng, nếu không muốn nói là sai. Khi số liệu không tuân theo luật phân phối chuẩn hay không đáp ứng các giả định trên, nhà nghiên cứu cần phải hóa chuyển số liệu để đáp ứng các giả định chung trên trước khi phân tích. Nếu số liệu không thể hoán chuyển, nhà nghiên cứu có thể áp dụng các phương pháp phân tích phi tham số (non-parametric method) như kiểm định Wilcoxon rank-sum test hay sử dụng phương pháp tái chọn mẫu (bootstrap), thay vì dùng phương pháp phân tích phương sai.

Thiếu sót 5: Không mô tả phương pháp dùng để phân tích sự khác biệt giữa hai nhóm trong phân tích phương sai.

Phương pháp phân tích phương sai được sử dụng để so sánh >2 nhóm. Nếu có 3 nhóm, chúng ta có thể so sánh nhóm 1 với 2, 1 với 3, và 2 với 3. Phân tích phương sai thường cung cấp hai kết quả quan trọng: giá trị kiểm định F và trị số *P*. Trị số *P* cho nhà nghiên cứu biết có ít nhất hai nhóm (trong các nhóm được phân tích) khác nhau có ý nghĩa thống kê, nhưng không cho biết cụ thể những nhóm nào! Để biết nhóm nào thật sự khác biệt, nhà nghiên cứu cần phải tiến hành bước thứ 2 trong qui trình phân tích: đó là phân tích *post hoc* – phân tích hậu kiểm. Có ít nhất là 5 phương pháp phân tích hậu kiểm, bao gồm Fisher's least significance test, Tukey, Student-Neuman-Keuls, Scheffe, Duncan, Bonferroni, v.v. Những phương pháp này có khi cho ra kết quả khác nhau vì khác nhau về giả định. Do đó, trong báo báo kết quả phân tích phương sai, nhà nghiên cứu phải trình bày rõ ràng phương pháp nào đã được áp dụng trong việc phát hiện những sự khác biệt và giả định đằng sau phương pháp phân tích.

Thiếu sót 6: Hiểu sai trị số P là xác suất của một giả thuyết khoa học.

Một hiểu lầm rất phổ biến là trị số P là xác suất giả thuyết vô hiệu. Chẳng hạn như nhà khoa học phát hiện mối liên quan giữa hút thuốc lá và ung thư với trị số $P = 0.04$, họ diễn

giải rằng xác suất *không* có mối liên quan là 4%. Suy ra, xác suất có mối liên quan giữa hút thuốc lá và ung thư phổi là 96%. Nhưng cách hiểu này sai. Trị số P không nói gì về xác suất của một giả thuyết khoa học. Trị số P chỉ giúp chúng ta bác bỏ giả thuyết vô hiệu, chứ không chứng minh giả thuyết nào cả.

Một hiểu lầm phổ biến khác cho rằng P là xác suất phát hiện sai. Ví dụ như nhà khoa học tính toán hệ số tương quan, và có kết quả r = 0.25, với P = 0.01, họ hiểu rằng xác suất kết quả này sai là 1%. Suy ra, xác suất kết quả đúng là 99%. Nhưng cách hiểu này hoàn toàn sai, vì trị số P không cho chúng ta biết là kết quả nghiên cứu đúng hay sai.

Thiếu sót 7: Trị số P là alpha (sai sót loại I).

Khi thiết kế một nghiên cứu khoa học (nhất là nghiên cứu lâm sàng), nhà khoa học phải xác định ngưỡng alpha và beta của nghiên cứu. Nói nôm na, alpha là dương tính giả (tức là xác suất mà nghiên cứu cho ra kết quả dương tính nhưng trong thực tế chẳng có liên quan gì). Còn beta là âm tính giả (tức là xác suất mà nghiên cứu cho ra kết quả âm tính, nhưng trong thực tế thì có liên quan). Do đó, có nhiều nhà khoa học hiểu rằng trị số P chính là alpha, nhưng cách hiểu đó sai. Sai vì hiểu lầm khái niệm kiểm định giả thuyết (test of hypothesis). Trị số P là kết quả của kiểm định thống kê (test of significance), chứ không phải kiểm định giả thuyết.

Thiết sót 8: Hiểu rằng trị số P càng thấp, mức độ ảnh hưởng càng cao.

Đây là một hiểu lầm rất tai hại. Nhiều nhà nghiên cứu so sánh trị số P để đánh giá mức độ ảnh hưởng. Chẳng hạn như họ tìm trong y văn và thấy ảnh hưởng của thuốc statin trong một nghiên cứu có trị số P = 0.01, còn nghiên cứu của họ có trị số P = 0.001, họ suy luận rằng mức độ ảnh hưởng họ quan sát cao hơn mức độ ảnh hưởng báo cáo trong y văn. Nhưng cách hiểu này sai, vì trị số P không phản ảnh mức độ ảnh hưởng so sánh giữa hai hay nhiều trị số P là không có ý nghĩa gì cả.

Thiếu sót 9: Kiểm định nhiều giả thuyết nhưng không hiệu chỉnh trị số P.

Phần lớn những nghiên cứu thực nghiệm báo cáo nhiều trị số *P*, vì nhà nghiên cứu kiểm định nhiều giả thuyết hay làm nhiều so sánh trong cùng một nghiên cứu, có khi cùng một dữ liệu. Chẳng hạn như nghiên cứu xác định hiệu quả của thuốc chống loãng xương có thể so sánh mật độ xương và hàng loạt marker chu chuyển xương giữa hai nhóm chứng và nhóm điều trị. Trong tình huống nhiều so sánh, xác suất kết quả dương tính giả (false positive) xảy ra rất cao. Nói theo ngôn ngữ thống kê, sai sót loại I (type I error) sẽ tăng nhanh với số lần kiểm định giả thuyết. Chẳng hạn như nếu chúng ta so sánh 15 biến giữa hai nhóm, và mỗi lần so sánh chúng ta chấp nhận sai sót loại I là 5% (a = 0.05), thì trong 15 so sánh đó, xác suất có ít nhất một so sánh sẽ có ý nghĩa thống kê (trị số *P* dưới 0.05) là $1 - (1 - 0.95)^{15} = 54\%$. Nói cách khác,

sai sót loại I bây giờ không phải là 5% nữa, mà là 54%! Do đó, nếu không điều chỉnh cho kiểm định nhiều giả thuyết (tiếng Anh là multiple comparison adjustment) thì chúng ta có thể đi đến kết luận sai, tức có thể phát hiện một sự khác biệt có ý nghĩa thống kê hoàn toàn do ngẫu nhiên chứ không phải do can thiệp.

Kiểm định nhiều giả thuyết hay nhiều so sánh xảy ra khi nhà nghiên cứu:

- Xác lập sự tương đương giữa nhóm bằng cách so sánh các biến số ban đầu (baseline variables) giữa hai nhóm trong một nghiên cứu lâm sàng đối chứng ngẫu nhiên, và họ hi vọng là sẽ không tìm ra một khác biệt nào (vì hai nhóm được chia nhóm ngẫu nhiên).

- So sánh giữa nhiều nhóm (pair-wise comparisons). Có nghiên cứu có nhiều hơn hai nhóm (chẳng hạn như 4 nhóm, A, B, C và D) và có khi nhà nghiên cứu muốn so sánh tất cả các nhóm A vs B, A vs C, A vs D, B vs C, v.v. Nói chung, số lần so sánh có thể lên đến $k(k-1)/2$, và kết quả dương tính giả hay tăng sai sót loại I lên rất cao khi có nhiều k và nhiều biến số so sánh.

- Kiểm định nhiều giả thuyết dựa trên nhiều biến số trong cùng một nghiên cứu.

- Phân tích thứ phát (secondary analysis) về mối tương quan giữa các biến trong nghiên cứu, dù những phân tích này không nằm trong dự tính lúc ban đầu.

- Chia nhóm và so sánh một cách tùy tiện. Trong nhiều trường hợp, nhà nghiên cứu có thể chia nhóm thành nam và nữ, độ tuổi thì có thể là 0-4, 5-9, 10-14, 15-19 nhưng cũng có thể 0-9, 10-19, và thậm chí 0-7, 8-12, 13-19, v.v. Khi có một biến liên tục thì có hàng vạn lần để chi biến đó thành những nhóm riêng lẻ, và trong trường hợp so sánh giữa các nhóm riêng lẻ như thế rất dễ dẫn đến kết quả sai.

- Phân tích lâm thời (interim analysis). Có nhiều nghiên cứu tuyển đối tượng theo thời gian, và cứ mỗi lần có thêm đối tượng, nhà nghiên cứu làm kiểm định giả thuyết thống kê. Do đó, những nghiên cứu theo thời gian, có khi nhà nghiên cứu phân tích rất nhiều lần, và trong số đó có những lần kết quả có ý nghĩa thống kê nhưng hoàn toàn do yếu tố ngẫu nhiên. Trong thực tế phân tích lâm thời thường được sử dụng khi nghiên cứu thử nghiệm lâm sàng ngẫu nhiên có nhóm chứng đã thu nhận được một số lượng đối tượng nghiên cứu nhất định (ví dụ 50% hay 75% cỡ mẫu dự tính). Mục tiêu là để có thể ngưng nghiên cứu sớm nếu đã có đủ bằng chứng về hiệu quả hay tác hại của biện pháp can thiệp (với tiêu chuẩn (stopping rules) được xác định trước trong đề cương nghiên cứu). Trong trường hợp này, giá trị P của phân tích cuối cùng

khi có đủ số liệu phải được hiệu chỉnh phù hợp nhằm tránh kết quả dương tính giả.

- So sánh giữa các nhóm cho nhiều thời điểm. Cũng có nghiên cứu mà trong đó nhà nghiên cứu theo dõi bệnh nhân ở nhiều thời điểm, nhà nghiên cứu kiểm định giả thuyết (hay so sánh) trong từng thời điểm. Đây cũng là một trường hợp *multiple comparison* thường hay thấy và những so sánh từng thời điểm như thế cũng làm tăng sai sót loại I.

Điều chỉnh trị số P trong trường hợp kiểm định nhiều giả thuyết thường là một yêu cầu, những cũng có khi không cần thiết. Do đó, nhà nghiên cứu cần phải biết lúc nào thì điều chỉnh và khi nào thì không.

Nếu một kiểm định ở độ ý nghĩa thống kê là a, xác suất sai sót loại I (tức xác xuất dương tính giả) được gọi là *comparisonwise error rate* (CER) a, có khi cũng gọi là *individual error rate*. Do đó, xác suất không bác bỏ giả thuyết là 1 – a. Nếu kiểm định *k* giả thuyết, xác suất không bác bỏ tất cả *k* giả thuyết vô hiệu nếu tất cả đều đúng là $(1 – a)^k$. Do đó, xác suất bác bỏ ít nhất là 1 giả thuyết vô hiệu được gọi là *experimentwise error rate (EER)* $= 1 – (1 – a)^k$. Xác suất này cũng được gọi là *global level* hay *familywise error rate*, vì *k* tests là một thí nghiệm. Nếu tất cả *k* giả thuyết độc lập, thì sẽ có khoảng *ka* kết quả có ý nghĩa thống kê nhưng trong thực tế là không có ý nghĩa thống kê. Nếu *k* giả thuyết không độc lập thì không có công thức đơn giản để ước tính số kết quả dương tính giả.

Nếu nhà nghiên cứu muốn kiểm soát CER thì không cần phải điều chỉnh trị số *P*. Ngoài ra, trong trường hợp phân tích khai thác (exploratory analysis) để tìm hiểu các mối liên hệ giữa các biến cũng không cần phải điều chỉnh cho trị số P. Tuy nhiên, nếu nhà nghiên cứu muốn kiểm soát EER thì cần phải điều chỉnh trị số *P*.
(Tham khảo: Bender R, Lange S. Adjusting for multiple testing – when and how? J Clin Epidemiol 2001;54:343-349).

Thiếu sót 10: Chỉ báo cáo kết quả qua trị số P.

Một bài báo y khoa viết như sau: "*The effect of the drug on lowering diastolic blood pressure was statistically significant (P<0.05).*" Ở đây, trị số P có thể là 0.049; tức có ý nghĩa thống kê (vì thấp hơn 0.05), nhưng rất gần với 0.05 mà có thể diễn giải tương tự như một trị số P bằng [chẳng hạn như] 0.051, tức không có ý nghĩa thống kê! Ngoài ra, , chúng ta không biết ảnh hưởng của thuốc trong việc hạ huyết áp là bao nhiêu, tức là chúng ta không biết ảnh hưởng của thuốc có ý nghĩa lâm sàng hay không.

Một nghiên cứu khác viết "*The mean diastolic blood pressure of the treatment group dropped from 110 to 92 mm Hg (P=0.02)*." Cách trình bày này tốt hơn cách trình bày trên, nhưng vẫn chưa đầy đủ. Giá trị trước và sau điều trị được báo cáo rõ ràng, nhưng không nói đến độ khác biệt. Thật ra, thuốc giảm huyết áp 18 mm Hg, và có ý nghĩa thống kê (P = 0.02), nhưng tác giả không cho chúng ta biết khoảng tin cậy 95% của độ khác biệt trước và sau điều trị.

Một cách viết tốt hơn nữa là "*The drug lowered diastolic blood pressure by a mean of 18 mm Hg, from 110 to 92 mm Hg (95% CI = 2 to 34 mm Hg; P=0.02)*." Ở đây, tác giả cho chúng ta biết ba thông tin quan trọng: huyết áp trước và sau điều trị; mức độ ảnh hưởng và khoảng tin cậy 95%; và trị số P. Khoảng tin cậy 95% có thể diễn giải nôm na rằng nếu thuốc được thử nghiệm trên 100 mẫu tương tự như nghiên cứu đang báo cáo, thì tính trung bình huyết áp giảm từ 2 đến 34 mm Hg trong 95 mẫu. Chúng ta biết rằng một giảm huyết áp A chỉ 2 mm Hg chẳng có ý nghĩa lâm sàng, nhưng giảm đến 34 mm Hg thì quả có ý nghĩa lâm sàng. Do đó, mặc dù huyết áp giảm trung bình là có ý nghĩa thống kê, mức độ khác biết có thể không phải lúc nào cũng có ý nghĩa lâm sàng; nói cách khác, kết quả nghiên cứu gần như khó kết luận. Để có kết luận dứt khoát, có lẽ chúng ta cần thêm bệnh nhân sao cho tất cả khoảng tin cậy 95% đều có ý nghĩa lâm sàng.

Thiếu sót 11: Diễn giải kết quả không có ý nghĩa thống kê như là một nghiên cứu negative.

Giả sử một nhà nghiên cứu so sánh huyết áp giữa hai nhóm, và kết quả không có ý nghĩa thống kê (statistically insignificant, P > 0.05). Nhà nghiên cứu phải quyết định sự không khác biệt đó có nghĩa là hai nhóm giống nhau (tương đương nhau), hay số liệu chưa đầy đủ để đi đến một kết luận chắc chắn hơn. Cần nói rằng một kết quả không có ý nghĩa thống kê không có nghĩa là hai nhóm giống nhau, mà chỉ có nghĩa là không thể bác bỏ giả thuyết vô hiệu, hay nghiên cứu này chưa đủ khả năng để bác bỏ giả thuyết vô hiệu. Giả thuyết vô hiệu (null hypothesis) là giả thuyết hai nhóm bằng nhau.

Nhiều nghiên cứu báo cáo kết quả không có ý nghĩa thống kê thường có power thấp, và do đó, không thể cung cấp câu trả lời dứt khoát. Nhà nghiên cứu có thể không "chứng minh" hai nhóm khác nhau, nhưng nhà nghiên cứu cũng không thể bác bỏ giả thuyết rằng hai nhóm có thể khác nhau. Người ta có câu *Absence of proof is not proof of absence hay Absence of evidence is not evidence of absence* (không có bằng chứng không có nghĩa là bằng chứng không có). Những nghiên cứu có power đầy đủ, một kết quả không có ý nghĩa thống kê có thể xem là một kết quả âm tính – negative (tức hai nhóm thật sự không khác nhau). Để "chứng minh" hai nhóm tương tự nhau thực tế cần phải tiến hành nghiên cứu tương đương (noninferior/equivalent) trong đó ngưỡng hiệu quả để đánh giá là tương đương phải được xác định ngay từ đầu.

Thiếu sót 12: Lẫn lộn giữa ý nghĩa thống kê (statistical significance) và ý nghĩa lâm sàng / ý nghĩa thực tế (clinical significance / practical significance).

Như đề cập ở đây, nhiều nhà nghiên cứu diễn giải một kết quả có ý nghĩa thông kê ($P < 0.05$) như là khẳng định có mối liên hệ sinh học hay có ý nghĩa lâm sàng. Thật ra, trị số *P* không có giá trị sinh học, và cũng không thể diễn giải như là có ý nghĩa sinh học hay ý nghĩa lâm sàng. Trong lâm sàng và sinh học, mức độ ảnh hưởng (effect size), mức độ khác biệt giữa hai hay nhiều nhóm mới là điều quan trọng. Có ý nghĩa thống kê là một điều kiện cần, nhưng chưa đủ để kết luận mối liên hệ hay ảnh hưởng là có thật.

Thiếu sót 13: Báo cáo trị số P không chính xác.

Có khá nhiều bài báo trình bày kết quả phân tích kèm theo trị số P được viết theo kiểu như "P < 0.05", "P < 0.01", hoặc "NS". Có lẽ "NS" ở đây có nghĩa là "not significant" hay "non-significance", tức không có ý nghĩa thống kê. Tất cả những cách báo cáo này là sai. Ngày xưa, vì không có phương tiện tính toán như bây giờ, nên người ta "lười biếng" viết như thế.

Còn ngày nay, chúng ta phải viết trị số P chính xác hơn. Thay vì viết P < 0.05, phải viết là P = 0.01. Nên nhớ mẫu tự P nên viết nghiêng. Ngoài ra, chỉ cần viết chính xác đến 3 số lẻ là đủ (ví dụ như P = 0.016 hoặc P < 0.001), chứ không cần viết quá rườm rà (như P = 0.0000012).

Thiếu sót 14: Khoảng tin cậy 95% là xác suất của kết quả.

Đây cũng là một hiểu lầm rất phổ biến trong khoa học. Tiêu biểu cho cách hiểu này là nhà nghiên cứu phân tích dữ liệu và có kết quả thuốc bisphosphonate giảm nguy cơ tử vong với nguy cơ tương đối 0.75, khoảng tin cậy 95% 0.35 đến 0.97; nhà nghiên cứu diễn giải rằng thuốc giảm nguy cơ tử vong 25%, và xác suất 95% là mức độ giảm dao động từ 3% đến 65%. Nhưng về mặt lí thuyết cách hiểu này sai. Khoảng tin cậy 95% không phải là xác suất 95%. Để tính được xác suất 95% đó (tức là xác xuất dao động của giá trị thật), phải dùng phương pháp Bayes.

Thiếu sót 15: Dùng trung bình và độ lệch chuẩn (SD) để mô tả một biến liên tục không tuân theo luật phân phối chuẩn.

Không như các biến định danh và định cấp vốn có thể mô tả bằng tần số (frequency) hoặc tỉ lệ (proportion) cho mỗi nhóm, các biến số liên tục có thể mô tả bằng một biểu đồ phân phối. Đối với các biến tuân theo luật phân phối chuẩn (normal distribution), có hai thông số chính là số trung bình và độ lệch chuẩn. Theo định nghĩa của luật

phân phối chuẩn, khoảng 67% các giá trị của nằm trong khoảng ±1 SD của số trung bình; khoảng 95% giá trị nằm trong khoảng ± 2 SD.

Tuy nhiên, nếu một biến không tuân theo luật phân phối chuẩn, thì số trung bình và độ lệch chuẩn sẽ không có ý nghĩa gì đáng kể. Đối với các biến không tuân theo luật phân phối chuẩn, các suy luận về 67% và 95% không còn đúng nữa. Trong trường hợp này, chúng ta nên dùng số trung vị (median) và số interquartile range để mô tả dữ liệu.

Phần lớn số liệu lâm sàng và sinh hóa không tuân theo luật phân phối chuẩn. Do đó, số trung vị và interquartile range nên được sử dụng thường xuyên hơn. Một cách tính nhẩm đáng tin cậy là nếu SD cao hơn phân nửa số trung bình (và số âm là số không khả dĩ về mặt sinh học) thì dữ liệu có lẽ không tuân theo luật phân phối chuẩn.

Thiếu sót 16: Dùng số trung bình và sai số chuẩn (standard error – SE) như là các chỉ số thống kê mô tả.

Số trung bình và độ lệch chuẩn (SD) là những chỉ số thống kê mô tả một mẫu nghiên cứu (study sample) với điều kiện biến số tuân theo luật phân phối chuẩn. Sai số chuẩn (standard error hay SE) là một chỉ số đo lường độ chính xác (precision) của một đặc điểm quần thể (population). Xin nhắc lại, SD áp dụng một mẫu nghiên cứu, SE áp dụng cho đặc điểm của một quần thể. SD phản ảnh độ dao động hay khác biệt giữa các cá nhân trong một mẫu nghiên cứu, còn SE phản ảnh độ dao động về một chỉ số như số trung bình giữa các mẫu tưởng tượng (vâng! tưởng tượng).

SE có thể ước tính từ SD bằng cách lấy SD chia cho căn số bậc hai của số cỡ mẫu. Do đó, SE lúc nào cũng thấp hơn SD. Nhiều nhà nghiên cứu không hiểu ý nghĩa của SE nên dùng nó như là một đo lường thay cho SD, và làm cho biến số có độ dao động thấp hơn so với thực tế. Một số nhà nghiên cứu sai lầm vì không hiểu (tức sai lầm có thể thông cảm), nhưng có những nhà nghiên cứu cao bồi cố tình lừa gạt người đọc bằng cách dùng SE thay cho SD và không nói rõ. Nói chung, nên dùng SD (chứ không phải SE) để mô tả một biến số.

Thiếu sót 17: Dùng "mean ± SD" không thích hợp.

Để mô tả một đại lượng liên tục, chúng ta hay dùng số trung bình (mean) và độ lệch chuẩn (standard deviation hay viết tắt là SD). Nhưng hai chỉ số này chỉ thích hợp cho các biến tuân theo luật phân bố chuẩn. Khi biến số không tuân theo luật phân bố chuẩn thì cách báo cáo mean ± SD được xem là một sai sót.

Ví dụ như có báo cáo viết như sau: "*The mean ± SD of testosterone was 0.92 ± 0.65 nmol/L*."

Có hai cái sai sót trong câu trên. Sai sót thứ nhất là biến số không tuân theo luật phân bố chuẩn, vì SD gần bằng với số trung bình. Chẳng lẽ dựa vào báo cáo trên, giá trị của testosterone có thể thấp đến -0.38 nmol/L? Vô lí. Cái sai thứ hai là không nên dùng dấu ± ở đây. Cách dùng đúng là dấu () chứ không phải ±.

Đối với biến không tuân theo luật phân bố chuẩn, cách mô tả thích hợp là trung vị (median) và bách phân vị 25% đến 75%. Ví dụ: "The median of total testosterone was 0.76 nmol/L, with interquartile range being from 0.15 to 1.12 nmol/L."

Thiếu sót 18: Dùng dấu nối khoảng tin cậy 95% không đúng qui ước.

Khoảng tin cậy 95% (KTC95%) có phần dưới và phần trên. Rất nhiều bài báo y học báo cáo hai phần này bằng cách dùng các dấu như gạch nối "-", dấu phẩy ",", thậm chí dấu chấm phẩy ";". Nhưng tất cả cách dùng đó đều không đúng qui ước. Dùng dấu "-" thường dễ gây hiểu lầm và nhập nhằng với dấu trừ. Dùng dấu phẩy thì có thể bị hiểu lầm số thập phân. Dùng dấu chấm phẩy thì sai hoàn toàn.

Cách dùng đúng theo qui ước là chữ "to". Một ví dụ tiêu biểu là viết như sau: "relative risk, 1.91; 95% confidence interval [CI], 1.75 to 2.09)." Chú ý dấu phẩy là viết sau relative risk, odds ratio, hazard ratio. Ngoài ra, chú ý trong cách viết chuẩn đó, dấu chấm phẩy là để thêm khoảng tin cậy 95%. Đây là cách viết chuẩn của Tập san *New England Journal of Medicine*.

Thiếu sót 19: Báo cáo các chỉ số thống kê hơn 2 số lẻ.

Thỉnh thoảng đọc báo cáo tôi thấy các đồng nghiệp hay trình bày những chỉ số như odds ratio (OR) và relative risk (RR) quá chính xác. Ví dụ như có báo cáo như sau: "BNP difference: OR = 0.998 (0.997–0.999)." Đây là một cách trình bày về độ chính xác ... không cần thiết. Thật ra, kết quả của ví dụ này có vẻ có vấn đề. Thường, chỉ cần hai số lẻ là đủ, như "OR, 0.76; 95% CI, 0.56 to 0.89" là đúng qui ước.

Thiếu sót 20: Chọn biến tiên lượng bằng phương pháp "stepwise" hoặc phân tích đơn biến có giá trị P < 0.05.

Một bài báo khoa học trên một tập san y học viết: "*Các biến có liên quan với tử vong trong phân tích đơn biến với mức ý nghĩa p<0.05 sẽ được đưa vào phân tích hồi qui đa biến logistic*". Nói cách khác, các tác giả tiến hành phân tích hai bước:

- Bước 1, phân tích từng biến một và lưu ý các biến có ý nghĩa thống kê (tức p < 0.05);
- Bước 2, cho tất cả các biến có ý nghĩa thống kê trong giai đoạn 1 vào một mô hình đa biến.

Đây là một sai lầm rất "vô tư" và khá phổ biến trong y văn và khoa học. Thậm chí, theo kinh nghiệm của người viết bài này, các nhà thống kê chuyên nghiệp cũng sai! Sai lầm này không hẳn là do tác giả cố ý, nhưng do hiểu lầm (hay chưa thông hiểu) cơ chế của các mô hình thống kê.

Vấn đề chính của cách chọn mô hình theo hai giai đọan trên là khi phân tích từng biến một (giai đoạn 1), mô hình hồi qui logistic không xem xét đến ảnh hưởng của các biến khác cùng một lúc. Chẳng hạn như nếu biến x_1 và x_2 có tương quan với nhau, thì phân tích giai đoạn 1 có thể chọn cả hai biến, nhưng trong mô hình đa biến (giai đoạn 2), có thể chỉ có x_1 có ý nghĩa thống kê, còn x_2 thì không (hay ngược lại), bởi vì thông tin của biến này đã hàm chứa trong thông tin của biến kia (do hai biến có liên quan nhau).

Một vấn đế khác, tinh vi hơn và "tế nhị" hơn, là ảnh hưởng của một biến trung gian, rất khó hay không thể kiểm soát trong giai đoạn 1. Trong trường hợp này, có thể hai biến có thể hai biến x_1 và x_5 (chẳng hạn) trong thực tế đều có ảnh hưởng đến Y, nhưng ảnh hưởng này chỉ tồn tại khi chúng xuất hiện bên nhau (cộng hưởng); do đó, khi phân tích riêng lẻ, chúng ta không phát hiện được ảnh hưởng của chúng, và do đó phân tích đơn giản trong giai đoạn 1 có thể bỏ qua cả hai biến!

Hiện nay, các phần mềm thống kê có sẵn một số thuật toán để phát hiện các biến độc lập cho mô hình đa biến, như thuật toán stepwise, backward, và forward. Nhưng ngay cả các thuật toán này, nhất là thuật toán stepwise và forward, vẫn có nhiều khiếm khuyết và cho ra những kết quả "dương tính giả", tức là những biến chẳng có liên quan gì đến biến phụ thuộc. Rất nhiều người không hiểu các thuật toán này nên vẫn áp dụng chúng một cách tùy tiện và hệ quả là có rất nhiều nghiên cứu với những kết quả sai trong khoa học.

Xây dựng một mô hình đa biến là một khoa học, nhưng cũng là một nghệ thuật. Khoa học tính liên quan đến các tiêu chuẩn định lượng và thuật toán thích hợp. Nghệ thuật tính liên quan đến những yếu tố có thể nói là chủ quan, đòi hỏi nhà nghiên cứu phải vận dụng kiến thức chuyên ngành để đi đến một mô hình có ý nghĩa lâm sàng. Mộ mô hình đa biến nếu chỉ thỏa mản các tiêu chuẩn khoa học vẫn chưa thể là một mô hình có ích. Một mô hình có ý nghĩa lâm sàng nhưng không đáp ứng các tiêu chuẩn khoa học không thể là một mô hình có độ tin cậy cao. Do đó, phân tích đa biến, dù là mô hình logistic hay hồi qui tuyến tính, là một phương pháp phức tạp, đòi hỏi nhiều thời gian để suy nghĩ và tính toán. Không thể và không nên để cho máy tính suy nghĩ dùm cho chúng ta.

Huyền thoại con số 30.

Một trong những "huyền thoại" khá phổ biến là nghiên cứu phải có cỡ mẫu trên 20 hay trên 30 mới có ý nghĩa vì với cỡ mẫu đó thì mới đạt yêu cầu của phân bố chuẩn. Thật ra, đây là một hiểu lầm tai hại! Phải khẳng định rằng không phải cứ cỡ mẫu 30 trở lên thì kết quả mới có ý nghĩa. Cỡ mẫu, như chúng ta đã xem qua trong Chương 18, tùy thuộc vào mô hình nghiên cứu, mức độ ảnh hưởng, độ dao động, và thể loại biến số. Có nghiên cứu chỉ cần cỡ mẫu dưới 20, nhưng cũng có nghiên cứu cần hơn 2000 đối tượng. Không có con số cố định tối thiểu và tối đa.

Tuy nhiên, con số này xuất phát từ một sự hiểu lầm, hay hiểu chưa đúng về phân bố thống kê và cỡ mẫu. Một số sách giáo khoa thường có một phát biểu chung chung về số cỡ mẫu "lớn" và "nhỏ". Chẳng hạn như cuốn *Probability and Statistical Inference* của Hogg và Tanis có viết rằng cỡ mẫu dưới 25 hay dưới 30 được xem là "nhỏ", và trên con số đó là "lớn". Nhưng ngưỡng nhỏ/lớn này không phải là cỡ mẫu cho nghiên cứu khoa học, mà là ngưỡng để tính xấp xỉ giữa phân bố chuẩn (normal distribution) và phân bố t (t distribution). Chúng ta biết rằng phân bố t là xấp xỉ phân phối chuẩn. Khi cỡ mẫu lớn, phân bố t và phân bố chuẩn gần như giống nhau. Ở đây, "lớn" có nghĩa là trên 30. Con số này, do đó, chẳng liên quan gì đến số cỡ mẫu cho nghiên cứu khoa học.

Có thể một hiểu lầm khác là liên quan đến Định lí giới hạn trung tâm (central limit theorem - CLT). Đại khái, CLT phát biểu rằng bất cứ chỉ số thống kê nào (như trung bình, phương sai, trung vị) đều tuân theo luật phân bố chuẩn hay gần với phân bố chuẩn nếu số cỡ mẫu đủ. "Đủ" ở đây thường được hiểu là trên 30. Cần nhấn mạnh rằng đây là một xấp xỉ về phân bố của chỉ số thống kê (statistic) như trung bình, trung vị, tỉ lệ, độ lệch chuẩn, v.v. chứ chẳng liên quan gì đến cỡ mẫu nghiên cứu.

***

Dĩ nhiên, những thiếu sót trên đây chưa kể đến những sai sót trong việc ứng dụng sai phương pháp, hoặc ứng dụng phương pháp đúng như sai giả định. Ngoài ra, còn có nhiều thiếu sót do kiến thức chưa được cập nhật. Trong thời gian khoảng 20 năm qua đã có khá nhiều tiến bộ đáng kể trong khoa học thống kê. Những tiến bộ này bao gồm, nhưng không giới hạn trong, phương pháp tái chọn mẫu bootstrap, mô hình hỗn hợp *mixed effects* hay *generalized estimating equation (GEE)*, phương pháp xử lí dữ liệu trống (missing values). Ngoài ra, phát triển về máy tính đã giúp phục hồi các phương pháp trong trường phái Bayes. Tất cả những phương pháp "mới" này giúp cho nhà khoa học rất rất nhiều và rất quan trọng. Không biết đến hay không dùng các phương pháp này có khi dẫn đến bỏ sót những khám phá quan trọng. Chẳng hạn như nhiều người được dạy rằng khi dữ liệu không tuân theo luật phân bố chuẩn hay vi phạm giả định thì phải dùng phương pháp phi tham số. Nhưng phương pháp phi tham số chỉ cho ra trị số P, mà chúng ta biết trị số P có rất nhiều vấn đề. Thay vì dùng phương pháp phi tham số, nhà khoa học có trình độ và hiểu biết sẽ dùng phương pháp bootstrap, vừa "lịch lãm" vừa cung cấp nhiều kết quả thú vị. Một sai sót phổ biến khác là dùng stepwise regression để chọn yếu

tố quan trọng. Có lẽ 95% các nhà khoa học đều dùng phương pháp stepwise, nhưng họ không biết rằng đó là một phương pháp rất dở, và cho ra nhiều sai sót (kết quả dương tính giả). Biết bao nhiêu "khám phá" sai trong khoa học có nguồn gốc từ phương pháp stepwise.

Trên đây là 20 thiếu sót (và sai sót và hiểu lầm) trong phân tích và báo cáo kết quả nghiên cứu có liên quan đến phân tích dữ liệu. Trong thực tế, sai sót và thiếu sót trong nghiên cứu khoa học, đặc biệt là phân tích thống kê rất phổ biến. Nên nhớ rằng khoảng 70% các bài báo khoa học bị từ chối là do phương pháp, trong đó gần phân nửa bị từ chối là do phân tích dữ liệu chưa đạt. Do đó, giảm các sai sót trên cũng là một cách nâng cao phẩm chất khoa học cho các nghiên cứu.