Diễn giải kết quả nghiên cứu khoa học

Nguyễn Văn Tuấn
Laboratory Head, Viện nghiên cứu y khoa Garvan
Giáo sư y khoa, Khoa Y, Đại học New South Wales
Giáo sư dịch tễ học, Khoa Y, Đại học Notre Dame Australia
Giáo sư y khoa tiên lượng, Đại học Công nghệ Sydney (UTS)
Australia

Một trong những khó khăn của người mới bắt đầu làm nghiên cứu khoa học là đọc và diễn giải một bài báo khoa học. Đối diện với một kết quả có ý nghĩa thống kê, câu hỏi kế tiếp là gì, và có thể tin vào kết quả này hay không, đó là những câu hỏi gai góc mà không phải lúc nào cũng có câu hỏi chính xác và đúng. Bài này sẽ điểm qua ba yếu tố quan trọng trong việc diễn giải kết quả nghiên cứu: yếu tố nhiễu, yếu tố bias, và yếu tố ngẫu nhiên.

Nhưng để bàn về 3 yếu tố đó, tôi muốn bàn qua về mục đích chung của các nghiên cứu khoa học. Nhìn chung và một cách tổng quát, nghiên cứu khoa học thường có 3 nhóm mục đích: phân loại, liên quan, và tiên lượng. Nhóm mục tiêu thứ nhất mang tính mô tả, như khám phá gen, phát hiện sinh vật mới, và nhà khoa học thường phân nhóm các phát hiện mới này. Nhóm mục tiêu thứ hai mang tính *association,* tức là tìm những qui luật chung về mối liên quan giữa các yếu tố. Chẳng hạn như nghiên cứu về mối liên quan giữa gan FTO và bệnh tiểu đường, nhà nghiên cứu qua phân tích dữ liệu có thể phát hiện những qui luật chung về sự khác biệt giữa các biến thể gen liên quan đến tỉ trọng mỡ trong cơ thể. Nhóm mục tiêu thứ ba là tiên lượng, hay *prediction*. Rất nhiều nghiên cứu khoa học có mục tiêu tìm các yếu tố quan trọng để phát triển mô hình tiên lượng. Những nghiên cứu về ung thư và một số bệnh mãn tính trong thực tế là nhằm phát hiện bệnh sớm hơn dựa vào "hồ sơ" môi trường và gen của mỗi cá nhân. Ba nhóm mục tiêu này định hình nhiều nghiên cứu khoa học trong thực tế.

Một cách nhìn khác là nguyên nhân và hệ quả. Thật vậy, đa số các nhà khoa học là những người đi tìm mối liên quan giữa nguyên nhân (cause) và hệ quả (effect). Chữ *cause* ở đây cần phải hiểu theo nghĩa rộng, bao gồm những yếu tố nói theo ngôn ngữ thống kê học là *independent variables* hay biến độc lập. Còn *effect* ở đây có thể hiểu là *outcome* trong y khoa. Nhưng một cách dễ hiểu và bao quát hơn là thay vì nói nguyên nhân và hệ quả, chúng ta có thể nói đến *exposure* và *outcome*. Như vậy, *exposure* ở đây phải hiểu là biến độc lập, là biến tiên lượng, là yếu tố nguy cơ; còn *outcome* là bệnh lí hay một đặc tính lâm sàng mà nghiên cứu quan tâm. Nếu nghiên cứu có mục tiêu tìm mối liên quan giữa uống cà phê và ung thư tụy, thì exposure ở đây là cà phê, và outcome là ung thư tụy. Đại đa số nghiên cứu y học là tìm hay xác định mối liên quan giữa exposure (E) và outcome (O).

Đối diện với một nghiên cứu mà kết quả cho thấy có mối liên quan giữa E và O, người đọc hay nhà khoa học làm gì để đánh giá mức độ khả tín của kết quả này? Một cách đơn giản nhất là xem xét đến mô hình nghiên cứu. Nếu mô hình nghiên cứu là RCT (randomized controlled trial) thì vấn đề đơn giản. Đối với nghiên cứu RCT so sánh hai nhóm (như nhóm dùng thuốc thật và nhóm dùng giả dược) được thiết kế tốt, tất cả các yếu tố liên quan đến O đều bằng nhau giữa hai nhóm bệnh nhân, nên khác biệt về O giữa hai nhóm chỉ có thể là do can thiệp. Nhưng đối với các nghiên cứu quan sát, nhà nghiên cứu không thể kiểm soát tất cả các yếu tố nguy cơ, nên có sự mất cân đối giữa hai nhóm so sánh (mắc bệnh và không mắc bệnh), nên vấn đề trở nên phức tạp hơn nghiên cứu RCT. Để diễn giải kết quả nghiên cứu quan sát một cách hợp lí, nhà nghiên cứu cần phải xem xét đến 3 yếu tố chính:

- Yếu tố ngẫu nhiên (chance hay random error)
- Yếu tố bias
- Yếu tố nhiễu (confounding effect)

Chỉ khi nào loại bỏ ba yếu tố này thì mối liên quan mới có thể xem là mối liên hệ nhân quả.

I. Yếu tố ngẫu nhiên

Bất cứ kết quả nghiên cứu nào cũng có yếu tố ngẫu nhiên. Nếu chúng ta quan sát mối liên quan giữa hút thuốc lá và ung thư phổi, thì câu hỏi đầu tiên đặt ra là: có phải đó là một quan sát ngẫu nhiên? Để trả lời câu hỏi này, các nhà nghiên cứu phải dùng phương pháp kiểm định thống kê (còn gọi là "test of significance") và trị số P. Nhưng vì trị số P thường hay bị hiểu sai, nên gần đây Hiệp hội Thống kê học Hoa Kì (American Statistical Association) mới ra một tuyên cáo về cách diễn giải trị số *P* (1).

1.1 Trị số P

Phương pháp kiểm định thống kê bắt đầu bằng một giả thuyết vô hiệu (null hypothesis), và đặt kết quả phân tích trong bối cảnh giả thuyết vô hiệu là đúng. Giả thuyết vô hiệu là một phát biểu nghịch đảo. Nếu giả thuyết là hút thuốc lá làm tăng nguy cơ ung thư phổi, thì giả thuyết vô hiệu phát biểu rằng nguy cơ bị ung thư phổi ở người hút thuốc lá và người không hút thuốc lá là như nhau. Do đó, phương pháp kiểm định thống kê được thực hiện như sau:

- Phát biểu giả thuyết vô hiệu H0 (không có khác biệt hay không có mối liên hệ);
- Làm nghiên cứu và thu thập dữ liệu (và gọi dữ liệu là D);

- Tính xác suất D xảy ra nếu H0 là đúng. Đây chính là trị số P;
- Nếu P ≤ 0.05, bác bỏ giả thuyết H0; nếu P > 0.05, chấp nhận giả thuyết H0.

Theo qui trình kiểm định thống kê như trên, trị số P có nghĩa là xác suất quan sát được dữ liệu D (hay dữ liệu cao/thấp hơn D) nếu giả thuyết vô hiệu là đúng. Nói theo ngôn ngữ xác suất ý nghĩa thật của trị số P là:

$$Pr(Data \mid H0)$$

Trị số P không cho chúng ta biết xác suất của giả thuyết H0 là bao nhiêu, đúng hay sai. Nhưng nhiều người hiểu lầm rằng trị số P là xác suất của giả thuyết vô hiệu! Không đúng. Trị số P là một xác suất có điều kiện (như công thức trên), và nó phản ảnh mức độ khả dĩ của dữ liệu (data) NẾU giả thuyết vô hiệu là đúng.

1.2 Vấn đề kiểm định nhiều giả thuyết

Một trong những vấn đề quan trọng trong diễn giải trị số P là vấn đề kiểm định nhiều giả thuyết. Một nghiên cứu thường kiểm định nhiều giả thuyết, chứ không phải đơn thuần chỉ một giả thuyết. Khi kiểm định nhiều giả thuyết thì xác suất tìm ra những mối liên hệ một cách ngẫu nhiên càng tăng cao.

Khi chúng ta kiểm định 1 giả thuyết, thường chúng ta chấp nhận sai sót 5% (còn gọi là sai sót loại I - type I error, còn gọi là α). Sai sót loại I cũng giống như dương tính giả, tức kết quả có ý nghĩa thống kê (P < 0.05) nhưng trong thực tế thì chẳng có mối liên hệ gì cả. Cũng có thể ví von sai sót loại I như kết quả chẩn đoán, tức là một cá nhân có kết quả xét nghiệm dương tính nhưng cá nhân đó thật ra không mắc bệnh. Xin nhắc lại, mỗi lần kiểm định giả thuyết, chúng ta chấp nhận sai sót 5%. Nói cách khác, xác suất đúng là 95% hay 0.95. Do đó, kiểm định 2 giả thuyết thì xác suất đúng cả hai là 0.95^2 = 0.9025; khi kiểm định 10 giả thuyết thì xác suất đúng cả 10 là 0.95^10 = 0.5987, tức khoảng 60%. Nói cách khác, khi kiểm định 10 giả thuyết, chúng ta có xác suất 40% tìm ra một kết quả có ý nghĩa thống kê, và kết quả này có thể chỉ là ngẫu nhiên. Nói chung, xác suất tìm ra một kết quả có ý nghĩa thống kê trong k test là 1 - (1 - α)^k. Khi k = 100 thì xác suất gần như 100% là chúng ta sẽ tìm ra một kết quả có ý nghĩa thống kê, nhưng kết quả này rất có thể là do yếu tố ngẫu nhiên.

Một vấn đề khác liên quan đến việc kiểm định nhiều giả thuyết là vấn đề "tra tấn dữ liệu" (data torture), còn gọi là "fishing expedition". Trong một nghiên cứu, thông thường nhà nghiên cứu kiểm định nhiều giả thuyết chính, nhưng ngoài giả thuyết còn phát sinh vấn đề chủ quan của nhà khoa học. Trong nhiều trường hợp, nhà khoa học muốn khai thác dữ liệu càng nhiều càng tốt (theo ý họ), chẳng hạn như phân tích khác biệt giữa nhóm can thiệp và nhóm chứng theo độ tuổi, theo giới tính, theo thành phần

kinh tế, hay nói chung theo những biến mà nhà khoa học nghĩ đến. Tình trạng "lang thang" với giả thuyết này có tên là "subgroup analysis" hay được ví von là "data torture", tức tra tấn dữ liệu cho đến khi có kết quả với trị số $P < 0.05$. Trên lí thuyết, các nghiên cứu với nhiều biến phân nhóm thì sẽ có vô vàn so sánh, và trong điều kiện đó chúng ta biết rằng xác suất tìm được một kết quả có ý nghĩa thống kê là gần như 100%, nhưng dĩ nhiên kết quả đó sai. Điều này nói lên rằng khi đọc một bài báo khoa học mà tác giả kiểm định quá nhiều giả thuyết khoa học và có kết quả có ý nghĩa thống kê (ví dụ như phát hiện mối liên hệ có ý nghĩa thống kê ở nhóm phụ nữ 45-54 tuổi sống ở thành phố và không hút thuốc không uống rượu) thì cần phải cẩn thận, vì rất có thể đó chỉ là một kết quả ngẫu nhiên.

1.3 Trị số P và cỡ mẫu

Trị số P trong thực tế là một kết quả hỗn hợp của hai thông tin: mức độ cao thấp của mối liên quan và độ chính xác của mối liên quan. Độ chính xác (precision) phụ thuộc vào cỡ mẫu; nghiên cứu với cỡ mẫu thấp thì độ chính xác thấp, nghiên cứu có cỡ mẫu cao thì độ chính xác cao. Nói cách khác, trị số P phản ảnh cùng một lúc cả hai thông tin: mức độ liên quan và cỡ mẫu.

Hai câu hỏi hay vấn đề đặt ra là: (i) nếu nghiên cứu có cỡ mẫu nhỏ (ví dụ $n = 10$) và có kết quả với trị số P cũng nhỏ (ví dụ $P = 0.01$); (ii) nếu nghiên cứu có cỡ mẫu lớn (ví dụ $n = 10000$) và trị số P cũng nhỏ (ví dụ $P = 0.001$), thì kết quả có đáng tin cậy không?

Chúng ta biết rằng nghiên cứu với cỡ mẫu nhỏ thì độ ảnh hưởng phải lớn mới có thể cho ra trị số P nhỏ. Nhưng một nghiên cứu với cỡ mẫu lớn hay rất lớn và dù độ ảnh hưởng thấp thì trị số P vẫn có thể nhỏ hay rất nhỏ. Như vậy, một nghiên cứu nhỏ mà cho ra kết quả với trị số P nhỏ thì đáng tin cậy? Không hẳn như thế. Câu chuyện liên quan đến James Heckman là một ví dụ thú vị về câu hỏi này:

James Heckman là một nhà kinh tế (Giải Nobel kinh tế, 2000) lí giải rằng Nhà nước nên cung cấp các dịch vụ "preschool" cho trẻ em. Bằng chứng cho lí giải này là những nghiên cứu nhỏ vào thập niên 1970. Trong một bài báo trên New York Times, ông phê bình những người phê bình ông về những nghiên cứu đó vì cỡ mẫu nhỏ. Ông lí giải rằng nghiên cứu với cỡ mẫu nhỏ mà cho ra kết quả với trị số P nhỏ thì bằng chứng càng mạnh (chứ không phải yếu):

"Also holding back progress are those who claim that Perry and ABC are experiments with samples too small to accurately predict widespread impact and return on investment. This is a nonsensical argument. Their relatively small sample sizes actually speak for — not against — the strength of their findings. ***Dramatic differences between treatment and control-group outcomes are usually not found in***

small sample experiments, *yet the differences in Perry and ABC are big and consistent in rigorous analyses of these data."*

Nhưng Heckman sai. Trong thực tế, những nghiên cứu với cỡ mẫu nhỏ có thể cho ra những kết quả rất lạc quan (trị số P thấp hay rất thấp, kiểu <0.001). Lí do là vì nghiên cứu với cỡ mẫu nhỏ thường bị sai số mẫu cao (sampling error) và có khi kết quả chỉ do vài con số ngoại vi.

Trong bảng dưới đây có 5 nghiên cứu với cỡ mẫu khác nhau; chỉ có 2 nghiên cứu mà tỉ số nguy cơ có ý nghĩa thống kê. Nhưng nghiên cứu A và C có kết quả đáng tin cậy hơn nghiên cứu B, D và E.

Nghiên cứu	Cỡ mẫu	Tỉ số nguy cơ	Trị số P
A	2500	1.4	**0.02**
B	500	1.7	0.10
C	2000	1.6	**0.04**
D	250	1.8	0.30
E	1000	1.6	0.06

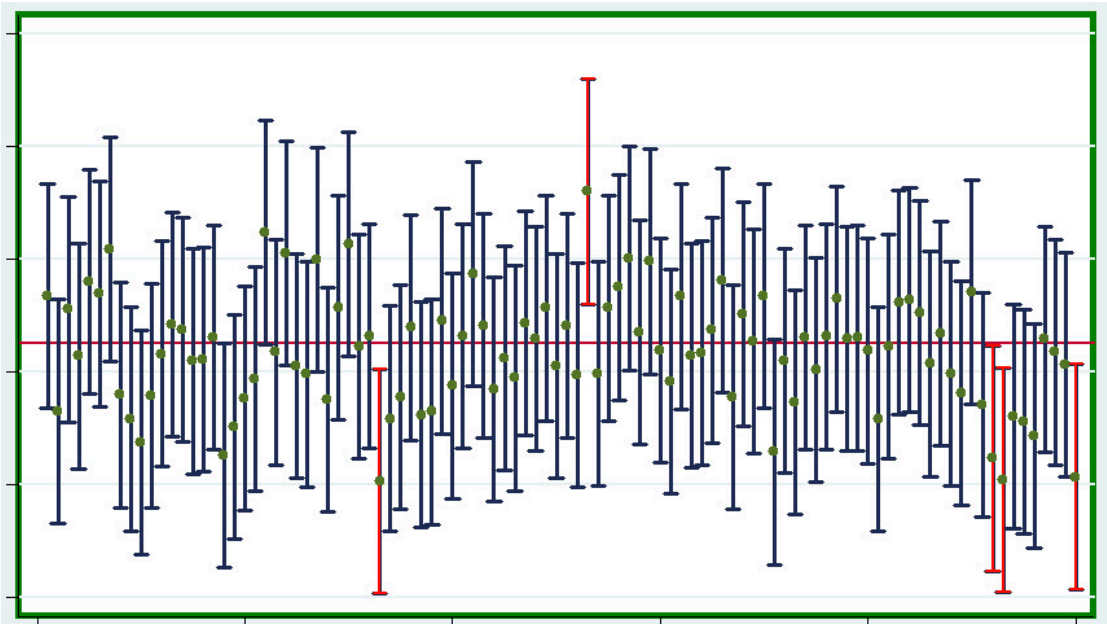
Nói chung. nghiên cứu với cỡ mẫu thấp ít khi nào có được trị số P thấp; nghiên cứu càng lớn càng có xu hướng tìm ra những kết quả với trị số P rất thấp. Do đó, một nghiên cứu với cỡ mẫu hàng triệu đối tượng mà có kết quả với trị số P = 0.07 thì kết quả đó không có gì quá ấn tượng!

1.4 Diễn giải khoảng tin cậy 95%

Những minh hoạ trên cũng có nghĩa là nếu một nghiên cứu chỉ được báo cáo bằng trị số *P*, thì chúng ta không thể nào xác định trị số P thấp là do mức độ liên quan lớn hay số cỡ mẫu lớn. Trong thực tế, một nghiên cứu với cỡ mẫu rất cao (hàng triệu đối tượng), nhưng độ ảnh hưởng rất thấp, thì trị số P vẫn có thể rất thấp. *Trị số P do đó không phản ảnh mức độ liên quan, không nói lên mức độ ảnh hưởng thấp hay cao*.

Vì trị số P không phản ảnh độ ảnh hưởng, nên giới nghiên cứu khoa học thường quan tâm đến khoảng tin cậy 95% (tiếng Anh là "*95% confidence interval*"). Khoảng tin cậy 95% có thể tính toán cho bất cứ chỉ số thống kê nào, như tỉ số odds, tỉ số nguy cơ, hệ số tương quan, v.v. Khoảng tin cậy 95% phản ảnh sự bất định của một mối liên quan, nhưng đồng thời cũng cung cấp thêm thông tin về sự khả dĩ của mối liên quan. Chẳng hạn như mối liên quan giữa gen A và bệnh tiểu đường có tỉ số odds là 5.2 (với KTC95 là 3.2 đến 7.2), chúng ta có thể hiểu rằng nếu nghiên cứu này được lặp lại rất nhiều lần, và mỗi lần với cùng cỡ mẫu, thì 95% các nghiên cứu đó sẽ có khoảng tin cậy 95% dao động trong khoảng 3.2 đến 7.2.

Nhưng trong thực tế thì có rất rất nhiều nhà khoa học hiểu sai KTC95. Họ thường hiểu rằng KTC95 là xác suất 95% là mức độ ảnh hưởng dao động trong khoảng đó. Chẳng hạn như với ví dụ trên, nhiều nhà khoa học hiểu rằng xác suất 95% là tỉ số odds dao động trong khoảng 3.2 đến 7.2. Nhưng cách hiểu này sai. Khoảng tin cậy 95% không phản ảnh xác suất của một sự ảnh hưởng.



Source: http://sites.nicholas.duke.edu/statsreview/ci/

Biểu đồ trên có thể dùng để giải thích ý nghĩa của khoảng tin cậy 95%. Có 100 nghiên cứu; mỗi nghiên cứu lấy mẫu từ quần thể, và ước tính KTC95 (dấu chấm màu xanh là trung bình, đường gạch đỏ ngang là tham số thật). Trong 100 nghiên cứu đó, có 5 nghiên cứu mà KTC95 lệch ra khỏi tham số thật; 95 nghiên cứu còn lại có KTC95 bao trùm tham số thật. Do đó, KTC95% có nghĩa là nếu nghiên cứu được lặp lại nhiều lần và mỗi lần có cùng cỡ mẫu, thì 95% các KTC95 sẽ dao động trong khoảng 3.2 đến 7.2 (ví dụ trên).

Khoảng tin cậy 95% còn cung cấp cho chúng ta một thông tin về sự bất định của độ ảnh hưởng. Bảng dưới đây trình bày kết quả nghiên cứu với khoảng tin cậy 95% cho tỉ số nguy cơ (risk ratio) cho 5 nghiên cứu với cỡ mẫu khác nhau. Chúng ta chú ý thấy nghiên cứu A và C có khoảng tin cậy 95% hẹp so với các nghiên cứu khác (B, E và E), và điều này có nghĩa là kết quả nghiên cứu A và C ổn định hơn các nghiên cứu khác.

Nghiên cứu	Cỡ mẫu	Tỉ số nguy cơ và khoảng tin cậy 95%	Trị số P
A	2500	1.4 (1.2 - 1.7)	**0.02**
B	500	1.7 (0.7 - 3.1)	0.10
C	2000	1.6 (1.2 - 2.1)	**0.04**
D	250	1.8 (0.6 - 3.9)	0.30
E	1000	1.6 (0.9 - 2.5)	0.06

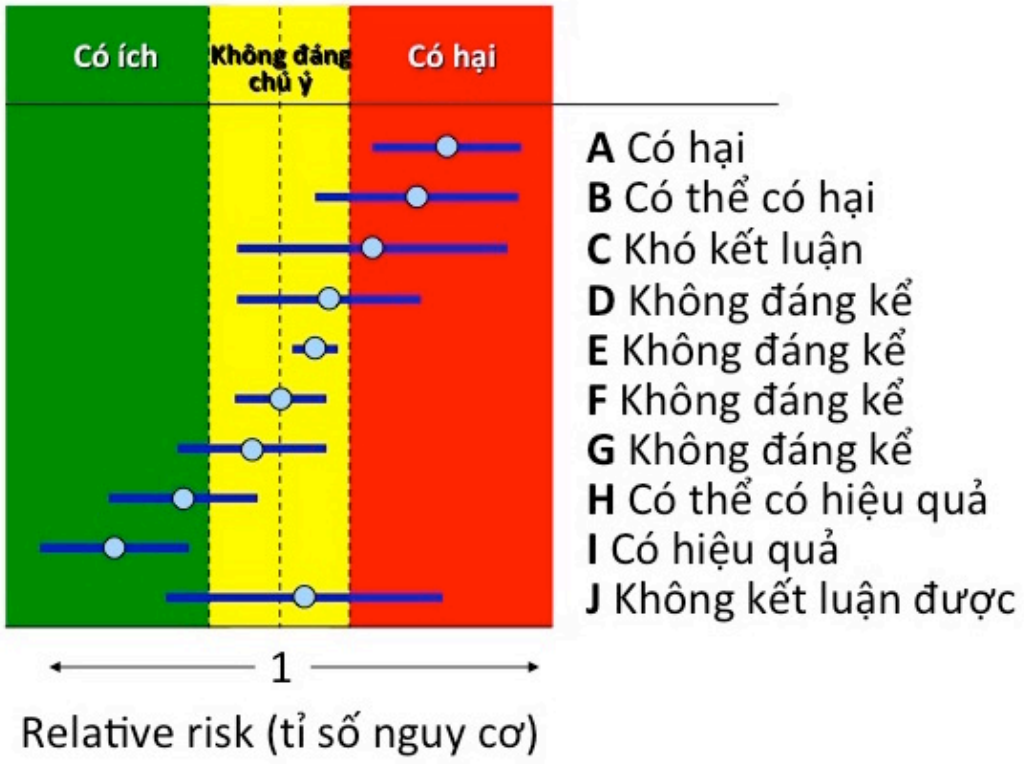
1.5 Diễn giải khoảng tin cậy 95% và ý nghĩa thực tế

Diễn giải kết quả nghiên cứu cần phải đặt trong ý nghĩa lâm sàng. Một kết quả có thể có ý nghĩa thống kê nhưng chẳng có ý nghĩa lâm sàng. Chẳng hạn như tỉ số nguy cơ gãy xương tay 1.05 có thể có trị số $P < 0.001$, nhưng mức độ tăng 5% nguy cơ thì không thể xem là có ý nghĩa lâm sàng. Do đó, trước khi đọc và diễn giải KTC95, nhà nghiên cứu cần phải xác định ngưỡng ảnh hưởng được xem là có ý nghĩa lâm sàng.

Chẳng hạn như nếu chúng ta xác định rằng nguy cơ đột quị tăng 10% trở lên được xem là có hại và giảm 10% trở lên được xem là có lợi, thì bất cứ yếu tố nguy cơ nào có tỉ số nguy cơ (RR) trong khoảng 1.01 đến 1.09 hay 0.91 đến 0.99 thì có thể xem là không có ý nghĩa lâm sàng. Nhưng nếu một yếu tố nguy cơ có RR từ 1.20 đến 1.70 thì rõ ràng là có hại. Tương tự, một can thiệp giảm nguy cơ đột quị với KTC95 RR từ (ví dụ) 0.5 đến 0.7 được xem là có lợi và có ý nghĩa lâm sàng.

Nhưng cũng có tình huống KTC95 dao động quá lớn, và kết quả này không thể cho chúng ta kết luận gì đáng tin cậy. Chẳng hạn như nếu tỉ số nguy cơ RR dao động từ 0.2 đến 3.5 thì kết quả này có thể phản ảnh cỡ mẫu của nghiên cứu còn nhỏ. Biểu đồ dưới đây cung cấp vài hướng dẫn để diễn giải một kết quả phân tích thống kê qua tỉ số nguy cơ. Vùng màu vàng có nghĩa là không có ý nghĩa lâm sàng; vùng màu xanh là có ích; và vùng màu đỏ là có hại.

Nghiên cứu A cho thấy KTC95 đều nằm trong vùng màu đỏ thì đó là một kết quả mang tính xác định (vừa có ý nghĩa lâm sàng, vừa có ý nghĩa thống kê). Nghiên cứu B tuy có ý nghĩa thống kê (vì phần dưới của KTC95 nằm ngoài đường tham chiếu), nhưng vì KTC95 bao trùm vùng không có ý nghĩa lâm sàng, nên chúng ta có thể kết luận rằng yếu tố này "có thể có hại". Nghiên cứu C thuộc vào nhóm "khó kết luận" vì KTC95 quá rộng và cũng không có ý nghĩa thống kê.



Nghiên cứu D tuy có KTC95 không rộng nhưng không có ý nghĩa thống kê và KTC95 cũng bao trùm vùng không có ý nghĩa lâm sàng, nên chúng ta có thể kết luận rằng yếu tố này có độ ảnh hưởng không đáng kể. Nghiên cứu E có KTC95 rất hẹp và có ý nghĩa thống kê, nhưng vì tất cả KTC95 nằm trong vùng không có ý nghĩa lâm sàng, nên chúng ta cũng có thể kết luận độ ảnh hưởng không đáng quan tâm. Cách diễn giải tương tự có thể áp dụng cho nghiên cứu F, G, H, I và J.

1.6 Bayes Factors

Bất cứ nghiên cứu khoa học nghiêm chỉnh nào cũng dựa trên một giả thuyết. Nhà khoa học rất muốn biết kết quả nghiên cứu có phù hợp với giả thuyết đặt ra hay không. Chẳng hạn như nếu chúng ta nghiên cứu về ảnh hưởng của thuốc điều trị loãng xương thì giả thuyết đặt ra có thể là tỉ lệ gãy xương ở nhóm điều trị thấp hơn tỉ lệ gãy xương ở nhóm không điều trị (nhóm chứng). Giả dụ như nếu chúng ta có kết quả với trị số P = 0.045. Câu hỏi đặt ra là trị số P đó có phản ánh giả thuyết chúng ta là đúng hay sai. Chúng ta biết rằng trị số P = 0.045 không có nghĩa rằng xác suất giả thuyết vô hiệu (thuốc không có hiệu quả) là 0.045. Hoàn toàn không có ý nghĩa đó. Trị số P = 0.045 có nghĩa là nếu thuốc không có hiệu quả, thì xác suất mà chúng ta quan sát được kết quả hiện tại là 4.5%. Như vậy, trị số P không trả lời câu hỏi xác suất giả thuyết thuốc có hiệu quả là bao nhiêu.

Để ước tính xác suất của một giả thuyết, chúng ta cần đến một phương pháp hoàn toàn khác: phương pháp Bayes. Đối với phương pháp Bayes, xác suất không phải là

một tần số mà phản ảnh "niềm tin" về một hiện tượng. Theo trường phái tần số, câu nói "xác suất ông Clinton đắc cử là 0.9" là vô nghĩa vì đó là một hiện tượng đơn lẻ (single event), nhưng đối với trường phái Bayes thì câu nói đó chấp nhận được vì nó phản ảnh độ tin tưởng của một cá nhân.

Để ước tính xác suất của một giả thuyết, phương pháp Bayes dựa vào hai thông tin: xác suất tiền định (prior probability) và số liệu thực tế (data) để ước tính xác suất hậu định (posterior probability). Định lí Bayes phát biểu rằng xác suất hậu định của một giả thuyết, P(H | data), là tích số của xác suất tiền định của giả thuyết đó, (PH), và xác suất dữ liệu nếu giả thuyết là đúng, P(data | H):

$$P(H \mid data) = P(H) \times P(data \mid H)$$

Như vậy, xác suất giả thuyết là đúng sau khi đã có dữ liệu phụ thuộc một phần vào giả thuyết tiền định, P(H).

Nhưng trong thực tế, chúng ta không biết hay rất khó xác định xác suất giả thuyết tiền định. Do đó, một cách khác để đánh giá sự khả tín của chứng cứ là tính hệ số Bayes, còn gọi là Bayes Factor (BF). BF là tỉ số của hai xác suất liên quan đến dữ liệu. Tử số là xác suất dữ liệu nếu giả thuyết vô hiệu là đúng, và mẫu số là xác suất dữ liệu nếu giả thuyết chính là đúng:

$$BF = \frac{P(data \mid H1)}{(Pdata \mid H0)}$$

Như vậy, BF là thước đo về chứng cứ. Nếu BF = 1, chứng cứ (dữ liệu) không nghiên về giả thuyết nào. Nếu BF < 1 thì dữ liệu nghiêng về giả thuyết vô hiệu. Ngược lại, nếu BF > 1 thì dữ liệu nghiêng về giả thuyết chính. BF nhìn chung là một thước đo tương đối khách quan vì nó không phụ thuộc vào xác suất của một giả thuyết nào. Một hướng dẫn chung về cách diễn giải BF là theo bảng dưới đây (tôi để nguyên tiếng Anh):

Bayes Factor	**Diễn giải**
>100	Decisive evidence for H1
30 đến 100	Very strong evidence for H1
10 đến 30	Strong evidence for H1
3 đến 10	Substantial evidence for H1
1 đến 3	Anecdotal evidence for H1
1	No evidence
1/3 đến 1	Anecdotal evidence for H0
1/10 đến 1/3	Substantial evidence for H0

1/30 đến 1/10	Strong evidence for H0
1/100 đến 1/30	Very strong evidence for H0
<1/100	Decisive evidence for H0

Hoá ra, dưới một số điều kiện phổ quát, BF có liên quan mật thiết với trị số P. Nói cách khác, nếu chúng ta có trị số P thì chúng ta cũng có thể ước tính BF. Trong một bài báo quan trọng công bố vào năm 1987 (2) và sau đó (3), Sellke và đồng nghiệp chỉ ra rằng BF tối thiểu có thể ước tính từ trị số P như sau:

$$\min BF = \frac{1}{-e \times P \times \log(P)}$$

hoặc minBF = exp(-0.5z^2). Trong đó, e = 2.718 là hằng số Euler, và z là chỉ số z. Chẳng hạn như một nghiên cứu với kết quả P = 0.045, chúng ta có thể ước tính BF như sau:

$$\min BF = \frac{1}{-2.718 \times 0.045 \times \log(0.045)} = 2.64$$

Như vậy, kết quả BF cho thấy bằng chứng cho giả thuyết H1 chỉ thuộc loại "anecdotal", tức rất yếu (dù P = 0.045 là có ý nghĩa thống kê).

Một mối liên hệ khác nữa cũng thú vị không kém là mối liên hệ giữa minBF và xác suất hậu định của giả thuyết nghiên cứu. Hoá ra, mối liên hệ này rất đơn giản:

$$\min post.p = \frac{1}{1 + \frac{1-q}{q \times \min BF}}$$

trong đó, q là xác suất tiền định của giả thuyết. Nếu chúng ta bắt đầu bằng q = 0.5 (tức 50%) thì với *min*BF = 2.64, xác suất hậu định tối thiểu của giả thuyết là:

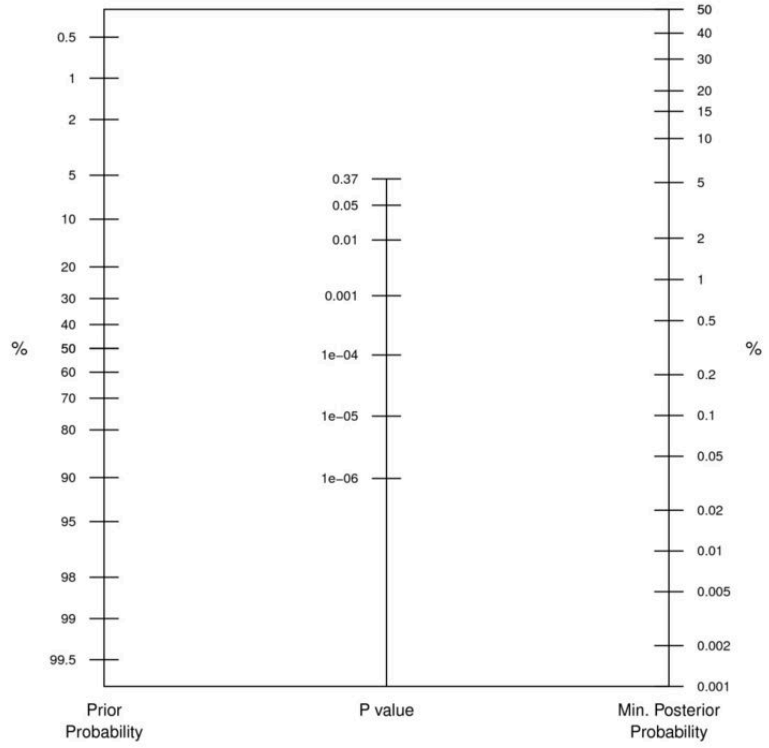
$$\min post.p = \frac{1}{1 + \frac{1-0.5}{0.5 \times 2.64}} = 0.72$$

Nói cách khác, cho dù P = 0.045 (tức có ý nghĩa thống kê), và với xác suất tiền định là 0.5, thì xác suất hậu định của giả thuyết chỉ 0.72. Nhưng nếu chúng ta không chắc chắn (trong trường hợp kiểm định nhiều giả thuyết), q có thể chỉ 5% (q = 0.05) thì xác suất hậu định chỉ 0.12. Bảng dưới đây trình bày xác suất hậu định cho các trị số P = 0.05, 0.01, 0.001, 0.0001 với xác suất tiền định 0.5.

Xác suất hậu định tối thiểu của giả thuyết nghiên cứu cho các trị số P và xác suất tiền định

Xác suất tiền định (q)	$P = 0.05$ (minBF = 2.46)	$P = 0.01$ (minBF = ~8.0)	$P = 0.001$ (minBF = 53.3)	$P = 0.0001$ (minBF = 399.5)
0.01	0.024	0.165	0.913	1.000
0.05	0.114	0.508	0.982	1.000
0.10	0.214	0.686	0.991	1.000
0.20	0.380	0.831	0.996	1.000
0.30	0.513	0.894	0.998	1.000
0.40	0.621	0.929	0.999	1.000
0.50	0.711	0.952	0.999	1.000

Trong thực tế, chúng ta có thể dùng nomogram do L. Held xây dựng để ước tính xác suất hậu định cho giả thuyết nghiên cứu (4). Cách dùng nomogram này rất đơn giản. Đầu tiên, đánh dấu vào trục số 1 về giả thuyết tiền định; sau đó đánh dấu vào cột trị số P; và vẽ một đường thẳng từ hai điểm đó đến trục thứ 3 và điểm giao chéo đó là xác suất hậu định của giả thuyết.



II. Yếu tố bias

Bias là chữ có nghĩa rất rộng trong nghiên cứu khoa học, và khó dịch sang tiếng Việt, nên tôi để nguyên chữ tiếng Anh. Trong nghiên cứu khoa học, bias được định nghĩa là một sai sót có hệ thống trong việc thiết kế và thực hiện nghiên cứu, và những sai sót này dẫn đến kết quả nghiên cứu bị lệch so với sự thật. Chẳng hạn như nếu trong thực tế (sự thật), người mắc bệnh tiểu đường có nguy cơ bị gãy xương tăng gấp 2 lần, nhưng trong nghiên cứu quan sát thì tỉ số nguy cơ là 4.5 lần, thì kết quả này được xem là *biased* (tính từ). Một ví dụ khác, nếu trong thực tế có mối liên quan giữa hút thuốc lá và ung thư phổi, nhưng nghiên cứu không phát hiện mối liên quan này, thì kết quả nghiên cứu cũng được xem là *biased.* Như vậy, bias ở đây có nghĩa là giá trị quan sát (observed value) không bằng với giá trị thật (true value).

Nói theo ngôn ngữ thống kê học, bias còn có nghĩa là giá trị của ước số (estimate) không bằng với giá trị của tham số (parameter). Chúng ta biết rằng parameter là giá trị của quần thể, còn estimate là giá trị của một mẫu, và chúng ta dùng estimate để ước tính giá trị khả dĩ của parameter. Nếu giá trị trung bình về chiều cao của 1000 phụ nữ là 165 cm, và nếu chúng ta biết rằng chiều cao trung bình của phụ nữa Việt Nam là 166 cm, thì sự khác biệt 1 cm đó được xem là *bias.* Bias trong thống kê học, do đó, là độ lệch giữa giá trị trung bình quan sát và giá trị trung bình trong quần thể.

Bởi vì kết quả của một nghiên cứu thường dựa vào một nhóm đối tượng, nên bias là một yếu tố có thể giải thích tại sao chúng ta quan sát mối liên hệ. Điều này cũng có nghĩa là khi diễn giải kết quả có ý nghĩa thống kê, chúng ta phải tự hỏi: *kết quả này có phải là do bias*? Có rất nhiều loại bias, nhưng tựu trung lại, có 3 loại bias chính cần phải quan tâm đến là *selection bias*, *information bias*, và *recall bias*. Selection bias có nghĩa là nghiên cứu chọn sai đối tượng. Information bias liên quan đến sai sót trong việc đo lường và xếp loại thông tin. Recall bias có nghĩa là những sai sót liên quan đến việc thu thập dữ liệu, mà thường đối tượng không nhớ những thông tin cá nhân một cách chính xác. Tất cả những bias này đều có thể dẫn đến kết luận sai của nghiên cứu.

2.1 "Selection bias" và câu chuyện cà phê và ung thư tuyến tụy

Brian MacMahon là một nhà dịch tễ học người Anh rất nổi tiếng, từng làm chủ nhiệm khoa dịch tễ học của Đại học Harvard trong một thời gian dài (1958 - 1988). Năm 1981, ông công bố một công trình nghiên cứu bệnh chứng về mối liên hệ giữa uống cà phê và ung thư tuyến tụy. Kết quả được công bố trên *New England Journal of Medicine* (5), nên thu hút rất nhiều sự chú ý của công chúng và các đồng nghiệp trong chuyên ngành. Trong nghiên cứu này, ông phỏng vấn 389 bệnh nhân ung thư tuyến tụy, và 644 người trong nhóm chứng (không bị ung thư tuyến tụy). Những thông tin

thu thập là hút thuốc lá, dùng bia rượu, trà, và cà phê. Ông báo cáo rằng có mối liên quan yếu giữa hút thuốc lá và nguy cơ ung thư, nhưng người dùng cà phê có nguy cơ ung thư gia tăng cao:

"There was a weak positive association between pancreatic cancer and cigarette smoking, but we found no association with use of cigars, pipe tobacco, alcoholic beverages, or tea. A strong association between coffee consumption and pancreatic cancer was evident in both sexes. [...] after adjustment for cigarette smoking, the relative risk associated with drinking up to two cups of coffee per day was 1.8 (95% confidence limits, 1.0 to 3.0), and that with three or more cups per day was 2.7 (1.6 to 4.7)."

Với kết quả này và với "niềm tin", ông khẳng định rằng đó là một mối liên quan thật. Và, từ niềm tin này, ông ngưng uống cà phê, và chỉ uống trà mà thôi!

Nhưng sau khi công trình nghiên cứu được công bố, nhiều nhóm nghiên cứu khác cũng làm nghiên cứu như thế, nhưng kết quả của họ không nhất quán với kết quả của MacMahon. Các nhóm nghiên cứu này không tìm thấy mối liên quan giữa uống cà phê và ung thư tuyến tụy.

Câu hỏi đặt ra là tại sao có sự khác biệt giữa kết quả của MacMahon và các nhóm sau đó? Sau khi phân tích kĩ qui trình tuyển chọn đối tượng nghiên cứu, các chuyên gia đi đến kết luận rằng MacMahon đã chọn ... sai đối tượng nghiên cứu trong nhóm chứng. Hóa ra, Giáo sư MacMahon chọn các đối tượng trong nhóm chứng từ bệnh nhân của các bác sĩ chuyên về các bệnh tiêu hóa. (Những bác sĩ này cũng là người chẩn đoán ung thư tuyến tụy.) Đứng về mặt nguyên tắc, MacMahon làm đúng, vì nhóm chứng nên chọn từ nguồn của nhóm bệnh. Nhưng vì các bệnh nhân của các bác sĩ chuyên khoa tiêu hóa thường [dĩ nhiên] mắc bệnh tiêu hóa, và họ không dùng cà phê hay được khuyến cáo không dùng cà phê.

	Nhóm bệnh	Nhóm chứng
Dùng cà phê	a	b
Không dùng cà phê	c	d

Do đó, nếu đặt trong bối cảnh nghiên cứu với bảng số liệu 2x2 (2 dòng và 2 cột) như trên, thì MacMahon có được d (số người không dùng cà phê trong nhóm chứng) nhiều hơn c. Do đó, khi tính odds odds1 = a/c và odds2 = b/d thì rõ ràng là odds1 cao hơn nhiều so với odds2. Do đó, kết quả của MacMahon thực chất thể hiện một sự bias. Bias này có tên là "selection bias".

2.2 Selection bias: câu chuyện HRT và bệnh tim mạch

Năm 1991, Stampfer và đồng nghiệp công bố một phân tích công trình nghiên cứu Nurses' Health Study (NHS) cho thấy thay thế hormone (HRT) có lợi cho sức khỏe phụ nữ sau mãn kinh (6). Công trình NHS theo dõi 48470 phụ nữ sau mãn kinh, tuổi từ 30 đến 63, những người này không có tiền sử bệnh tim mạch. Trong thời gian 10 năm theo dõi, họ phụ nữ dùng HRT có nguy cơ mắc bệnh xơ vữa động mạch (coronary heart disease, CHD) thấp hơn nhóm không dùng HRT (tỉ số nguy cơ [RR] 0.56, khoảng tin cậy 95% dao động từ 0.40 đến 0.80). Sau đó, một phân tích tổng hợp do Grady et al (7) thực hiện cho thấy HRT giảm CHD với RR 0.65. Tác giả kết luận rằng *"current estrogen use is associated with a reduction in the incidence of coronary heart disease as well as in mortality from cardiovascular disease."* Nhưng đây là những kết quả của nghiên cứu quan sát.

Đến cuối thập niên 1990s, một công trình nghiên cứu RCT có tên là HERS (Heart and Estrogen-progestin Replacement Study) trên tập san JAMA (8). Kết quả của HERS cho thấy HRT không làm giảm CHD, nhưng làm tăng nguy cơ đột quị và ung thư vú. Năm 2002, một nghiên RCT khác có tên là WHI (Women Health Initiative Study) cũng được công bố trên JAMA. Kết quả của WHI cũng cho thấy HRT làm tăng CHD, tăng nguy cơ đột quị, tăng nguy cơ ung thư vú, nhưng giảm nguy cơ gãy cổ xương đùi (9).

Tỉ số nguy cơ liên quan đến HRT và các bệnh lí mãn tính

Outcome	**HERS**	**WHI**
CHD	0.99 (0.80 - 1.22)	1.29 (1.02 - 1.65)
Đột quị	1.23 (0.89 - 1.70)	1.41 (1.07 - 1.85)
Pulmonary embolism	2.79 (0.89 - 8.73)	2.13 (1.39 - 3.25)
Ung thư vú	1.30 (0.77 - 2.19)	1.26 (1.00 - 1.59)
Hip fracture	1.10 (0.49 - 2.50)	0.66 (0.45 - 0.98)
Tử vong	1.08 (0.84 - 1.38)	0.98 (0.82 - 1.18)
Global index	NA	1.15 (1.03 - 1.28)

Tại sao có sự khác biệt lớn về ảnh hưởng của HRT giữa nghiên cứu quan sát và nghiên cứu RCT? Lí do chính là do nghiên cứu quan sát không thể ngẫu nhiên hóa các đối tượng nghiên cứu. Do đó, hai nhóm (nhóm dùng HRT và nhóm không dùng HRT) rất khác nhau về các yếu tố quan trọng có thể có liên quan đến bệnh tim mạch. Chẳng hạn như nhóm dùng HRT hóa ra là những phụ nữ thuộc thành phần kinh tế cao, học thức cao, thu nhập cao, hay luyện tập thể dục, ít hút thuốc lá, ít dùng bia rượu, họ hay quan tâm đến sức khỏe, hay nói chung là khỏe mạnh hơn nhóm không dùng HRT. Khi các yếu tố này được phân tích một cách đứng đắn, các nhà nghiên cứu không tìm thấy sự khác biệt về nguy cơ bệnh tim mạch giữa hai nhóm! Do đó, kết quả

các nghiên cứu quan sát là bias, và bias này có tên là "*healthy bias*". Healthy bias là một dạng của selection bias.

2.3 "Recall bias" và câu chuyện dị tật bẩm sinh

Một trong những ca nghiên cứu nổi tiếng hay dùng làm minh họa cho vấn đề recall bias là nghiên cứu bệnh chứng về yếu tố nguy cơ liên quan đến dị tật bẩm sinh. Hầu hết các nghiên cứu đều tìm ra một yếu tố nào đó. Nhưng mối liên quan mà các nghiên cứu này phát hiện có thể là do bias trong việc thu thập thông tin. Các cha mẹ có con dị tật bẩm sinh thường có động cơ tìm hiểu các yếu tố nguy cơ, và họ thường hay nhìn lại quá khứ phơi nhiễm của mình. Nói cách khác nhìn theo bảng số liệu dưới đây, nhà nghiên cứu sẽ có số *a* (số người có yếu tố nguy cơ trong nhóm bệnh) cao hơn nhóm chứng. Do đó, odds1 = a/c cao hơn odds2 = b/d.

	Nhóm bệnh	Nhóm chứng
Phơi nhiễm	a	b
Không phơi nhiễm	c	d

Điều này cũng có nghĩa là tỉ số odds (odds ratio) mà nhà nghiên cứu quan sát có thể chỉ là do thông tin thu thập dẫn đến sự mất cân đối về yếu tố nguy cơ giữa hai nhóm. Tình trạng này còn gọi là *differential bias*, *reporting bias*, hay *rumination bias*.

Bias là một vấn đề lớn trong các nghiên cứu quan sát. Nhiều nhà nghiên cứu đã thống kê được 256 loại bias, nhưng trong thực tế tôi nghĩ con số còn cao hơn, nhất là trong thời đại "Dữ liệu Lớn". Một trong những sai sót "sáng chói" trong kết luận liên quan đến dữ liệu Lớn là nghiên cứu của Google trong việc dựa vào các tweets để biết sự chuyển biến của trận bão Sandy (2016). Tính theo số tweets, Google thấy rất nhiều tweets từ Manhattan và rất ít từ các vùng như Breezy Point, Coney Island, và Rockaway, và họ kết luận rằng Manhattan đang bị cơn bão Sandy. Nhưng thật ra đây là một dạng bias về thông tin, vì những vùng bị ảnh hưởng bởi cơn bão mất điện nên người dân không dùng điện thoại di động được. Ngoài ra, Manhattan là vùng nhà giàu nên số người có điện thoại di động cũng cao hơn các vùng khác. Bài học ở đây là nếu chỉ dựa vào con số mà không xem xét đến bối cảnh thì rất dễ dẫn đến sai lầm mà không có dữ liệu, dù lớn hay nhỏ, có thể giải quyết. Trong bài này, tôi chỉ bàn đến 3 loại bias phổ biến trong nghiên cứu quan sát, nhưng cũng đủ để cảnh báo tất cả chúng ta phải cẩn thận trong việc diễn giải một kết quả có ý nghĩa thống kê.

III. Yếu tố nhiễu (confounding effects)

Chữ *confounder* xuất phát từ tiếng Latin *confundere* có nghĩa là "pha trộn vào nhau". Do đó, định nghĩa của *confounding* theo Rothman là ảnh hưởng của yếu tố phơi nhiễm bị lẫn lộn với ảnh hưởng của một yếu tố khác và dẫn đến bias. Trong thực tế, một yếu tố được xem là confounder (tạm kí hiệu là Z), nếu yếu tố đó hội đủ 3 điều kiện sau đây:

- Z có liên quan với outcome (Y);
- Z có liên quan đến exposure (X); và
- Z không nằm trong cơ chế liên quan giữa X và Y.

3.1 Yếu tố nhiễu: dẫn nhập

Giả dụ như chúng ta nghiên cứu về mối liên quan giữa bệnh loãng xương (osteoporosis, OS) và thoái hóa khớp (osteoarthritis, OA) bằng mô hình nghiên cứu đoàn hệ. Giả thuyết đặt ra là loãng xương là yếu tố nguy cơ gây bệnh thoái hóa khớp. Giả dụ như qua 5 năm theo dõi một nhóm gồm 1000 bệnh nhân loãng xương và 1000 người không loãng xương, ghi nhận kết quả như sau:

	OA	Không OA	Tỉ lệ OA
Loãng xương	380	620	0.38
Không loãng xương	110	890	0.11

Kết quả trên cho thấy loãng xương có liên quan đến tăng nguy cơ OA. Tỉ số nguy cơ là 0.38 / 0.11 = 3.5, có nghĩa là nguy cơ bị OA ở bệnh nhân loãng xương cao gấp 3.5 lần so với nguy cơ OA ở nhóm không loãng xương.

Nhưng kết quả này có thể là do chúng ta chưa xem xét đến yếu tố độ tuổi. Chúng ta biết rằng người bị loãng xương thường là cao tuổi (trên 60). Chúng ta cũng biết rằng những bệnh nhân thoái hóa khớp cũng là bệnh nhân cao tuổi. Do đó, một cách kiểm tra có sự ảnh hưởng của yếu tố nhiễu (độ tuổi) hay không, chúng ta thử phân tích theo hai nhóm tuổi: nhóm dưới 45-79 và nhóm 80 tuổi trở lên.

Nhóm 45-79 tuổi

	OA	Không OA	Tỉ lệ OA
Loãng xương	20	80	0.20
Không loãng xương	90	810	0.01

Nhóm 80+ tuổi

	OA	Không OA	Tỉ lệ OA
Loãng xương	360	540	0.40
Không loãng xương	20	80	0.20

Chúng ta thấy ở nhóm 45-79 tuổi, tỉ số nguy cơ RR = 0.20 / 0.10 = 2. Ở nhóm 80+ tuổi, tỉ số nguy cơ cũng bằng 2 (0.4 / 0.2). Như vậy, ở đây tỉ số nguy cơ thật là 2.0, chứ không phải 3.5 như phân tích ở trên. Lí do của sự khác biệt về kết quả là vì mối liên quan giữa loãng xương và thoái hóa khớp chịu sự ảnh hưởng của yếu tố độ tuổi.

3.2 Confounding by indication

Trong các nghiên cứu quan sát về mối liên quan giữa dùng thuốc và bệnh lí, yếu tố nhiễu có tên là "confounding by indication" (yếu tố *nhiễu vì chỉ định*) là một cơn ... ác mộng. Một trong những trường hợp tiêu biểu là mối liên quan giữa thuốc chống trầm cảm và sa sút trí thuệ (cognitive deficit). Kết quả của nghiên cứu cho thấy những người dùng thuốc chống trầm cảm có nguy cơ sa sút trí tuệ tăng cao. Nhưng đây là một kết quả có thể do nhiễu vì chỉ định. Lí do đơn giản là vì người dùng thuốc này chắc chắn có tiền sử trầm cảm hay đang bị trầm cảm, và trầm cảm thì có liên quan đến sa sút trí tuệ; do đó, mối liên quan giữa thuốc chống trầm cảm và sa sút trí tuệ có thể là ảo.

Một nghiên cứu nổi tiếng công bố trên *New England Journal of Medicine* về aspirin và nhồi máu cơ tim (MI) là một ca tiêu biểu cho yếu tố nhiễu vì chỉ định (10). Nghiên cứu thoạt đầu được thiết kế theo mô hình RCT, với hai nhóm bệnh nhân: nhóm aspirin được cho uống aspirin (325 g/ngày) và nhóm giả dược. Kết quả cho thấy bệnh nhân nhóm aspirin có nguy cơ MI giảm rõ rệt so với nhóm chứng, và do đó, ủy ban y đức quyết định ngưng công trình nghiên cứu.

Nhưng sau khi công trình nghiên cứu bị ra lệnh dừng lại, tất cả bệnh nhân của 2 nhóm aspirin và nhóm chứng được cho phép uống aspirin, và họ được theo dõi 7 năm. Kết quả của nghiên cứu quan sát này được công bố trên *Arch Int Medicine* vào năm 2000 (11). Kết quả cho thấy aspirin có liên quan đến giảm nguy cơ MI 38% (tỉ số nguy cơ RR 0.72; khoảng tin cậy 95%, 0.55 đến 0.95), không có liên quan đến đột quị, nhưng giảm nguy cơ tử vong (RR 0.64; khoảng tin cậy 95%, 0.54 đến 0.77). Bảng dưới đây so sánh kết quả RCT và nghiên cứu quan sát:

Tỉ số nguy cơ (risk ratio) liên quan đến nhồi máu cơ tim, đột quị và tử vong

Outcome	**RCT**	**Nghiên cứu quan sát**
Nhồi máu cơ tim	0.56 (0.45 - 0.70)	0.72 (0.55 - 0.95)
Đột quị	2.14 (0.96 - 4.77)	1.02 (0.74 - 1.39)
Tử vong	0.96 (0.60 - 1.54)	0.64 (0.54 - 0.77)

Nhưng kết quả còn thú vị hơn trên. Những bệnh nhân uống aspirin lại là những người từng có tiền sử gia đình với MI, cao huyết áp, dùng bổ sung vitamin E, và vài yếu tố khác như cân nặng, bia rượu, và luyện tập thể dục. Người uống aspirin do đó thường cao tuổi hơn, hay hút thuốc lá, dùng bia rượu, quá cân và béo phì, và đã có nguy cơ MI cao. Chính vì những yếu tố nguy cơ này mà họ chọn dùng aspirin. Do đó, mối liên quan mà nhà nghiên cứu quan sát giữa aspirin và MI thật ra chỉ là một ảnh hưởng không hẳn của aspirin mà là do các yếu tố nguy cơ.

3.3 Nghịch lí Simpson (Simpson Paradox)

Một trong những hình thức của ảnh hưởng nghịch đảo là "Nghịch lí Simpson", được nhà thống kê học Edward Simpson mô tả lần đầu vào năm 1951 (tuy nhiên hiện tượng này đã được ghi nhận trước đó khá lâu) (12). Nghịch lí Simpson có thể định nghĩa như là một sự ảnh hưởng không nằm trong kì vọng khi tổng hợp nhiều nhóm nhỏ. Nếu chúng ta quan sát ảnh hưởng của một can thiệp trong 3 nhóm A, B, C một cách nhất quán, Nghịch lí Simpson tiên đoán rằng nếu chúng ta tổng hợp dữ liệu của 3 nhóm thành một quần thể, thì ảnh hưởng của can thiệp trong quần thể có thể ngược lại với ảnh hưởng trong từng nhóm!

Bảng dưới đây là một dữ liệu thật trích từ một nghiên cứu dịch tễ học. Các nhà nghiên cứu quan sát tỉ lệ tử vong ở nam và nữ của hai nhóm bệnh nhân phân nhóm theo giới tính. Kết quả cho thấy ở nam, tỉ lệ tử vong ở nhóm can thiệp thấp hơn nhóm chứng ở cả nam (40% so với 50%) và nữ (5% so với 15%).

Nhóm	**Nam**		**Nữ**	
	Tử vong	Sống	Tử vong	Sống
Can thiệp	80 (40%)	120	5 (5%)	95
Chứng	20 (50%)	20	39 (15%)	221

Tuy nhiên, khi các nhà nghiên cứu tổng hợp nam và nữ thành một nhóm như bảng số liệu dưới đây:

Nhóm	**Nam + Nữ**	
	Tử vong	Sống
Can thiệp	85 (**28%**)	215
Chứng	59 (**20%**)	241

thì tỉ lệ tử vong trong nhóm can thiệp lại *cao hơn* nhóm chứng (28% so với 20%).

Hiện tượng này xảy ra, vì sự phân bố nhóm (hay yếu tố nguy cơ) mất cân đối giữa các nhóm đối tượng. Chẳng hạn như chúng ta thấy ở nam, số đối tượng trong nhóm can thiệp cao gấp 5 lần nhóm chứng, còn ở nữ, số đối tượng thuộc nhóm can thiệp thì rất thấp (100) so với nhóm chứng (260). Trong trường hợp này, phân tích bằng cách gộp 2 nhóm nam và nữ có thể cho ra kết quả sai. Tuy nhiên, nếu nhà nghiên cứu có kinh nghiệm, họ có thể sử dụng phương pháp phân tích thống kê để hiệu chỉnh cho sự mất cân đối nhóm.

Hiện tượng Nghịch lí Simpson thường hay thấy trong nghiên cứu quan sát vì cách phân nhóm một cách "tự nhiên". Điều này dẫn đến sự mất cân bằng trong phân nhóm và kết quả chung có thể bị nhiễu / sai. Do đó, diễn giải kết quả của nghiên cứu quan sát, nhất là kết quả tổng hợp, cần phải hết sức cẩn thận vì có thể chịu sụ chi phối của Nghịch lí Simpson.

3.4 "Nghịch lí môi sinh" (ecologic fallacy)

Émile Durkheim (1858 - 1917) là một nhà xã hội học người Pháp rất nổi tiếng với những công trình nghiên cứu về mối liên quan giữa tín ngưỡng và tự tử. Trong cuốn sách nổi tiếng *"Le Suicide"* xuất bản năm 1897, ông trình bày số liệu cho thấy tỉ lệ tự tử ở những địa phương có nhiều tín đồ đạo Tin Lành (protestants) cao hơn những địa phương có nhiều tín độ đạo Công Giáo (Catholics). Ông kết luận rằng do nền tảng xã hội và sự kiểm soát xã hội ở những địa phương có đông tín độ Công Giáo, nên nguy cơ tự tử ở người Công Giáo thường thấp.

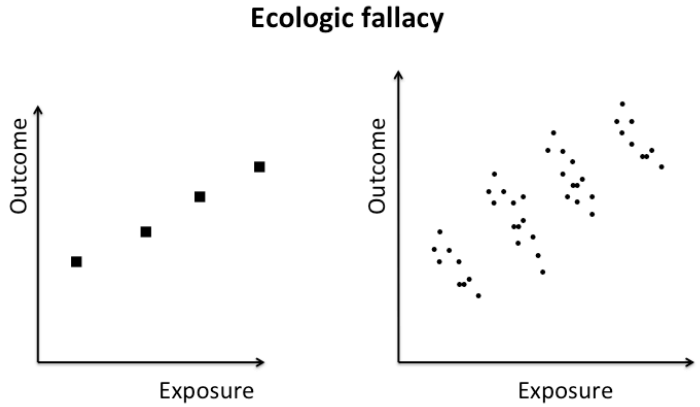
Nhưng đó là một sai sót quan trọng. Sai sót này có tên là nghịch lí mô sinh hay *ecologic fallacy*. Đây là một nghịch lí rất phổ biến trong khoa học xã hội và kinh tế học. Không có địa phương nào là có 100% tín đồ Tin Lành hay 100% tín đồ Công Giáo. Những địa phương có nhiều tín đồ Công Giáo thì cũng có tín đồ Tin Lành nhưng họ là thiểu số. Rất có thể vì cộng đồng Tin Lành là thiểu số trong cộng đồng Công Giáo đa số làm cho tín đồ Tin Lành tự tử nhiều hơn.

Nhưng sai lầm lớn nhất của Durkheim là đơn vị phân tích của ông là địa phương, chứ không phải cá nhân. Vấn đề chúng ta quan tâm là tín đồ Tin Lành có nguy cơ tự tử

cao hơn tín đồ Công Giáo. Nhưng Durkheim là dùng dữ liệu của địa phương, chứ không phải dữ liệu ở cấp độ cá nhân. Dữ liệu cấp địa phương hay cộng đồng không thể dùng để suy luận cho mối liên hệ ở cấp cá nhân.

Để hiểu vấn đề hơn, chúng ta có thể xem qua 2 biểu đồ dưới đây. Biểu đồ bên trái cho thấy có mối liên quan dương tính giữa exposure và outcome; khi giá trị của exposure càng cao thì giá trị outcome cũng càng cao. Dựa vào kết quả này, nhà nghiên cứu có lẽ kết luận rằng mối liên hệ giữa exposure và outcome là khá chặt chẽ.

Nhưng thật ra 4 điểm trong biểu đồ này là số trung bình của 4 địa phương (xem biểu đồ bên phải). Biểu đồ bên phải cho thấy ***ở mỗi địa phương***, mối liên hệ giữa exposure và outcome là nghịch đảo! Ở mỗi địa phương, khi giá trị của exposure tăng thì giá trị của outcome giảm. Như vậy, mối liên hệ thật giữa exposure và outcome là nghịch đảo. Biểu đồ bên trái "che giấu" mối liên hệ thật, vì nó dựa vào số liệu tổng kết cho từng địa phương, và khi tính hệ số tương quan trên cơ sở địa phương rất dễ dẫn đến sai lầm.



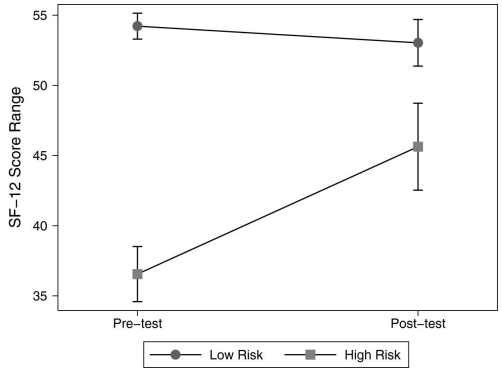
3.5 Hiệu ứng hồi qui về số trung bình

Một trong những wisdom của người Việt tôi thấy rất hay là câu nói "*Không ai giàu ba họ, chẳng ai khó ba đời.*" Nhìn trong thực tế, chúng ta thấy câu này rất ứng nghiệm với những nhà giàu nứt vách như gia đình của Công tử Bạc Liêu. Khi ông giàu có tột đỉnh, ông dám chơi ngông bằng cách đốt tiền để nấu chè (theo huyền thoại, chứ có thật hay không cũng chẳng rõ), nhưng vài đời sau thì con cháu của ông sống trong cảnh túng quẫn. Ngược lại, cũng có những gia đình xuất thân rất nghèo vài thế hệ, nhưng sau đó thì có người làm giàu nứt vách. Chúng ta thấy sau 1975, có những gia đình xưa kia rất nghèo khó, chủ yếu làm nghề nông hay theo du kích, nhưng sau đó thì giàu có nhất nhì nước. Nhưng nếu có dịp định lượng, tôi nghĩ con cháu của những gia đình rất giàu hay những gia đình rất nghèo tính trung bình là những người có xu hướng có thu nhập trung bình. Đây cũng là một hiện tượng hay thấy trong khoa học mà tôi tạm dịch là *"Hiệu ứng hồi qui về số trung bình"*.

Tiếng Anh gọi hiệu ứng này là "*regression toward the mean effect",* và nó có nguồn gốc từ nhà khoa học lừng danh Francis Galton. Trong bài báo "Regression towards Mediocrity in Hereditary Stature," Galton quan sát thấy có mối tương quan giữa chiều cao của con và chiều cao trung bình của cha và mẹ (ông gọi là mid-parent height). Khi chiều cao của cha và mẹ ở số trung bình trong quần thể, thì chiều cao của con bằng chiều cao trung bình đó (68.2 in); khi chiều cao trung bình của cha mẹ cao hơn số trung bình thì chiều cao của con có xu hướng thấp hơn cha mẹ và quay về số trung bình; và khi chiều cao của cha mẹ thấp hơn trung bình thì chiều cao của con cao hơn cha mẹ nhưng cũng có xu hướng quay về số trung bình quần thể. Galton gọi qui luật hay hiện tượng này là "regression toward the mean" (RTM).

Qui luật hồi qui về số trung bình (RTM) còn thấy trong hầu hết các thí nghiệm khoa học tự nhiên và khoa học xã hội. Trong các nghiên cứu trước - sau (before-after study), bệnh nhân được đo lường hai lần: trước khi điều trị (hay can thiệp) và tạm kí hiệu là X0, và sau khi điều trị, tạm gọi là X1. Do đó, mỗi bệnh nhân, chúng ta có thể tính hiệu quả bằng hiệu số D = X1 - X0. Hiệu ứng RTM tiên đoán rằng D sẽ liên quan nghịch đảo với X0.

Giả dụ rằng chúng ta làm một nghiên cứu can thiệp một nhóm bệnh nhân bằng tập thể dục và outcome là mật độ xương (MĐX). Mật độ xương của một nhóm bệnh nhân được đo trước và sau khi tập thể dục. Hiệu ứng RTM cho biết những bệnh nhân có MĐX cao hay rất cao sẽ có mức độ giảm cao sau khi can thiệp. Nhưng RTM cũng nói rằng những bệnh nhân có MĐX thấp hay rất thấp, thì sau khi can thiệp MĐX sẽ tăng. Xu hướng tăng và giảm đó quay về số trung bình. Do đó, khi chúng ta quan sát MĐX tăng thì điều đó không hẳn là do tập thể dục mà có thể là do hiệu ứng RTM. Tương tự, khi chúng ta quan sát những người có MĐX cao lúc ban đầu, sau khi tập thể dục, MĐX của họ suy giảm, thì điều đó không hẳn có nghĩa là tập thể dục làm giảm sức mạnh của xương. Đó chỉ là do hiệu ứng của RTM.



Ví dụ về hiệu ứng RTM: Nhóm với score ban đầu thấp sau khi can thiệp thì score tăng; nhóm có score cao lúc ban đầu, sau khi can thiệp có xu hướng giảm.

Nguồn: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-119

Hiệu ứng RTM có thể thấy trong rất nhiều bối cảnh. Từ nghiên cứu giáo dục (điểm học sinh) đến nghiên cứu tâm lí, tất cả đều chịu sự tác động của hiệu ứng RTM. Những học sinh thoạt đầu có điểm thấp và sau này có điểm cao, hay những học sinh ban đầu có điểm cao sau này có điểm thấp, có thể là do hiệu ứng RTM. Trong phẫu thuật, những bệnh nhân nặng nhất thường có hiệu quả cao nhất, nhưng đó có thể là do hiệu ứng RTM chứ không phải do phẫu thuật.

Để giảm thiểu ảnh hưởng của hiệu ứng RTM, các nhà nghiên cứu có kinh nghiệm thường thiết kế nghiên cứu có nhóm chứng (tức không can thiệp). Đó cũng chính là lí do tại sao các nghiên cứu RCT thường có nhóm chứng. Không có nhóm chứng, những "hiệu quả" mà chúng ta quan sát rất khó diễn giải, và có lẽ chẳng có hiệu quả gì cả!

IV. Chín điều kiện để đánh giá một mối liên hệ nhân quả

Bradford Hill là một nhà thống kê học nổi tiếng trong thế giới y khoa. Ông là người đầu tiên thực hiện công trình nghiên cứu RCT về thuốc kháng sinh và bệnh lao phổi. Năm 1965, ông công bố một luận văn (bài báo) được nhiều trích dẫn sau đó (13). Trong bài báo, ông đề ra 9 tiêu chí, hay nói đúng hơn là 9 điều kiện, để đi đến một kết luận về nguyên nhân và hệ quả (tiếng Anh):

- Strength of association
- Consistency
- Specificity
- Temporality
- Biological gradient
- Plausibility
- Coherence
- Experiment
- Analogy

4.1 Mức độ liên quan (strength of association)

Những mối liên nhân quả thật sự thường có mức độ liên quan cao. Múc độ liên quan ở đây thể hiện qua các chỉ số thống kê như tỉ số nguy cơ, tỉ số odds, hiệu số, hay mức độ ảnh hưởng nói chung (effect size). Chẳng hạn như trong mối liên hệ giữa loãng xương và gãy cổ xương đùi có tỉ số nguy cơ bằng 3 hoặc cao hơn (bệnh nhân loãng xương có nguy cơ gãy xương cao gấp 3 lần so với người không bị loãng xương). Đó là một mối liên hệ khá cao, và là một tín hiệu của mối liên hệ nhân quả. Nhưng cũng có những mối liên hệ thấp với tỉ số odds cỡ 1.05 hay thậm chí 1.5, và những kết quả này cần phải đặt trong vòng nghi ngờ.

4.2 Tính nhất quán (consistency)

Một mối liên hệ nhân quả thật sự thường được quan sát ở nhiều đối tượng khác nhau, môi trường khác nhau, địa phương khác nhau, và ở nhiều thời điểm khác nhau. Chẳng hạn như mối liên hệ giữa hút thuốc lá và ung thư phổi, hay giữa loãng xương và gãy xương đã được lặp lại ở nhiều đối tượng tại nhiều quần thể ở gần như bất cứ thời điểm nào. Do đó, các mối liên hệ này rất có thể là liên hệ nhân quả, chứ không phải ngẫu nhiên hay do một yếu tố khác.

4.3 Tính đặc hiệu (specificity)

Khái niệm đặc hiệu được phát kiến vào thời các bệnh truyền nhiễm còn phổ biến. Khái niệm đặc hiệu có nghĩa là một nguyên nhân chỉ dẫn đến một bệnh (như vi trùng lao Mycobacterium tuberculosis chỉ là nguyên nhân của bệnh lao). Nhưng khái niệm này không còn thích hợp cho thời đại của các bệnh mãn tính phức tạp vốn chịu ảnh hưởng của nhiều yếu tố. Chẳng hạn như hút thuốc lá không chỉ là nguyên nhân của ung thư phổi, mà còn ảnh hưởng đến nhiều bệnh mãn tính khác như xơ vữa động mạch và tiểu đường. Do đó, tuy đặc hiệu là một điều kiện cho mối liên hệ nhân quả, nhưng không phải là điều kiện bắt buộc.

4.4 Tính thời gian (temporality)

Đây là điều kiện liên quan đến thời gian. Điều kiện này có nghĩa là yếu tố nguy cơ hay nguyên nhân phải xảy ra trước khi bệnh xảy ra. Chẳng hạn như nếu loãng xương là nguyên nhân của gãy xương, thì loãng xương phải xảy ra trước biến cố gãy xương. Điều này cũng có nghĩa là các nghiên cứu đoàn hệ và nghiên cứu RCT mới có thể phát biểu về các mối liên hệ nguyên nhân và hệ quả.

4.5 Mối liên hệ liêu lượng và ảnh hưởng (biologic gradient)

Theo Hill, một mối liên hệ nhân quả thường thể hiện qua sự liên hệ mang tính liều lượng (dose - response relationship). Chẳng hạn như mối liên hệ giữa hút thuốc lá và ung thư phổi thì điều kiện này có nghĩa là người hút thuốc lá càng nhiều thì nguy cơ mắc bệnh ung thư càng cao. Tương tự, người bị suy giảm mật độ xương càng nhiều thì nguy cơ bị gãy xương càng tăng cao. Trong thực tế cũng có những mối liên hệ không tuân theo mối tương quan tuyến tính, mà có thể là phi tuyến tính, tức là chỉ ở một ngưỡng nào đó của yếu tố phơi nhiễm thì nguy cơ bệnh mới tăng nhanh. Nếu một mối liên hệ không tuân theo qui luật dose - response thì rất có thể đó không phải là mối liên hệ nhân quả.

4.6 Cơ sở sinh học (plausibility)

Một mối liên hệ nhân quả phải có cơ sở sinh học. Chẳng hạn như mối liên hệ giữa mật độ xương và gãy xương có thể giải thích rằng vì mật độ xương có liên quan trực tiếp với sức mạnh của xương, và do đó người có mật độ xương thấp thì xương rất yếu, khó có thể chịu được một ngoại lực. Tuy nhiên, trong thực tế, có rất nhiều mối liên hệ không thể giải thích bằng cơ sở sinh học, bởi vì đơn giản là chúng ta chưa có mô hình sinh học để giải thích. Do đó, cơ sở sinh học cũng không phải là điều kiện nhất thiết cho suy luận về mối liên hệ nhân quả.

4.7 Tính khúc chiết (coherence)

Điều kiện khúc chiết có nghĩa là mối liên hệ nhân quả thường không mâu thuẫn với những hiểu biết về quá trình phát sinh của bệnh. Chẳng hạn như mối liên hệ giữa hút thuốc lá và ung thư phổi là khúc chiết vì nó nhất quán với nhiều quan sát khác như (i) khi tỉ lệ người hút thuốc lá càng tăng theo thời gian thì số ca ung thư cũng tăng theo; (ii) những nước có nhiều người hút thuốc lá cũng là những nước có nhiều ca ung thư phổi; (iii) bằng chứng từ histology cho thấy bronchial epithelium của người hút thuốc lá khác với người không hút thuốc lá; và (iv) bằng chứng từ nghiên cứu trên chuột cũng cho thấy chuột phơi nhiễm khói thuốc lá có nguy cơ bị ung thư phổi cao. Tất cả những chứng cứ này nhất quán với nhau, và đó là tín hiệu cho thấy mối liên hệ giữa hút thuốc lá và ung thư phổi là mối liên hệ nhân quả.

4.8 Chứng cứ từ thí nghiệm (experiment)

Theo John Stuart Mill, nếu A là nguyên nhân của B, thì nếu tất cả các yếu tố khác cố định, một sự thay đổi A sẽ dẫn đến thay đổi B. Và, thí nghiệm có kiểm soát (controlled experiment) là một phương pháp để kiểm định giả thuyết về mối liên hệ nhân quả giữa A và B. Trong mối liên hệ giữa loãng xương và gãy xương, nếu can thiệp tăng mật độ xương dẫn đến giảm nguy cơ gãy xương, thì đó là một tín hiệu cho mối liên hệ nhân quả.

4.9 Sự tương đương (analogy)

Theo Hill, nhà nghiên cứu phải dùng các ví von giữa các mối liên quan để giải thích (thuyết phục) một mối liên hệ nhân quả. Chẳng hạn như mối liên hệ giữa thuốc thalidomide và rubell, chúng ta có thể chấp nhận một chứng cứ tương tự giữa một loại thuốc với một bệnh truyền nhiễm ở phụ nữ đang mang thai. Tuy nhiên, trong thực tế ít ai nghĩ đến điều kiện này vì nó vừa định tính vừa chủ quan, thiếu tính khoa học!

***

Như đề cập ở phần đầu, nếu nghiên cứu được thiết kế theo mô hình RCT thì cách diễn giải kết quả rất đơn giản. Do các yếu tố nhiễu và bias đều được kiểm soát chặt chẽ qua ngẫu nhiên hoá, nên sự khác biệt giữa hai nhóm nghiên cứu (ví dụ như nhóm can thiệp và nhóm chứng) có thể xem như là một mối liên hệ nhân quả. Nhưng đối với phần lớn các nghiên cứu quan sát (như mô hình nghiên cứu bệnh chứng, cắt ngang, đoàn hệ) thì chúng ta không thể kiểm soát hết tất cả các yếu tố nhiễu và bias, do đó vấn đề diễn giải kết quả trở nên phức tạp.

Khi đối diện với một kết quả nghiên cứu có ý nghĩa thống kê ($P < 0.05$), nhà nghiên cứu phải suy nghĩ đến 3 cách giải thích. Thứ nhất, kết quả đó có thể là ngẫu nhiên, và nghiên cứu chỉ "may mắn" có được, chứ không phản ảnh mối quan hệ trong thực tế. Thứ hai, kết quả đó có thể là do sự ảnh hưởng của các yếu tố bias, kể cả bias trong

việc chọn đối tượng nghiên cứu, bias trong đo lường và thu thập thông tin. Thứ ba, kết quả đó cũng có thể do các yếu tố nhiễu. Các yếu tố nhiễu có thể là có một biến thứ ba mà nhà nghiên cứu không biết hay không kiểm soát được, có thể là bias vì chỉ định (đặc biệt trong các nghiên cứu về thuốc), có thể là nghịch lí Simpson, nghịch lí môi sinh, hay hiệu ứng hồi qui về số trung bình.

Khi 3 yếu tố ngẫu nhiêu, bias và nhiễu được loại bỏ thì vấn đề kế tiếp là câu hỏi: đó có phải là một mối liên hệ nhân quả. Tôi đã trình bày 9 tiêu chuẩn (hay đúng ra là điều kiện) của Bradford Hill để diễn giải một mối liên hệ nhân quả. Không phải bất cứ mối liên hệ nhân quả nào cũng đáp ứng 9 điều kiện này, nhưng đó là những điểm để tham khảo trong việc diễn giải một kết quả nghiên cứu khoa học.

Tham khảo và đọc thêm:

(1) Ronald L. Wasserstein và Nicole A. Lazar. The ASA's Statement on *p*-Values: Context, Process, and Purpose. The American Statistician 2016;70:129-133.

(2) Berger JO, Sellke T: Testing a point null hypothesis: Irreconcilability of *P* values and evidence (with discussion). J Am Stat Assoc. 1987, 82: 112-139.

(3) Sellke T, Bayarri MJ, Berger JO: Calibration of *p* Values for Testing Precise Null Hypotheses. Am Stat. 2001, 55: 62-71.

(4) L Held. A nomogram for *P* values. BMC Medical Research Methodology 201010:21.

(5) MacMahon B, et al. Coffee and cancer of the pancreas. N Engl J Med. 1981 Mar 12;304(11):630-3.

(6) Stampfer M et al. Postmenopausal estrogen therapy and cardiovascular disease. Ten-year follow-up from the nurses' health study. N Engl J Med. 1991 Sep 12;325(11):756-62.

(7) Grady D, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med. 1992 Dec 15;117(12):1016-37.

(8) Hulley S, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/Progestin Replacement Study (HERS) Research Group. *JAMA* 1998;280:605–13.

(9) Rossouw JE, et al. Risks and benefits of estrogen plus progestin in healthy ostmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321–33.

(10) Final report on the aspirin component of the ongoing Physicians' Health Study. Steering Committee of the Physicians' Health Study Research Group. N Engl J Med. 1989 Jul 20;321(3):129-35.

(11) Cook NR, Hebert PR, Manson JE, Buring JE, Hennekens CH. Self-selected posttrial aspirin use and subsequent cardiovascular disease and mortality in the physicians' health study. Arch Intern Med. 2000;160(7):921-8.

(12) Simpson EH. "The Interpretation of Interaction in Contingency Tables". Journal of the Royal Statistical Society, Series B 1951;13: 238–241.

(13) Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.