Lâm sàng thống kê 22

Đo lường ảnh hưởng: Odds ratio, relative risk, risk ratio, hazard ratio

Nguyễn Văn Tuấn

Trong bài "Lâm sàng thống kê 14", tôi đã giải thích sự liên hệ và khác biệt giữa *odds ratio* (OR) và *relative risk* (RR), và những khó khăn trong việc diễn giải OR. Gần đây có một số bạn hỏi tôi về *risk ratio* và *hazard ratio* (HR) là gì, và cách sử dụng cũng như diễn giải các chỉ số này như thế nào. Trong bài này, tôi sẽ giải thích ngắn gọn các thuật ngữ này. Tôi sẽ không dịch sang tiếng Việt, vì có ý để các bạn đọc biết thuật ngữ nguyên thủy tiếng Anh để khi ra ngoài có dịp "hội nhập" cùng các đồng nghiệp quốc tế.

Nhưng trước khi vào đề, tôi muốn kể cho các bạn một câu chuyện trong lịch sử y khoa có liên quan đến nguy cơ và xác suất. Christiaan Barnard là nhà giải phẫu đầu tiên ghép tim (heart transplantation) trên thế giới. Trong hồi kí của mình, ông thuật lại một câu chuyện về bệnh nhân thay tim đầu tiên trên thế giới của Barnard: đó là Louis Washkansky. Khi được đưa vào phòng giải phẫu, Washkansky đang say mê đọc sách trên giường như không để ý gì đến một sự kiện lịch sử y khoa sắp xảy ra. Barnard vào phòng giải phẫu, tự giới thiệu với Washkansky, và giải thích tường tận rằng ông sẽ cắt bỏ trái tim của Washkansky và thay vào đó là một trái tim mới lành mạnh hơn. Barnard nói thêm: ông sẽ có *cơ may* bình phục (there is a *chance* that you can get back to normal life again). Washkansky không hỏi cái cơ may đó là bao nhiêu, không hỏi ông có thể sống bao lâu nữa, mà chỉ nói "tôi sẵn sàng" và quay lại tiếp tục đọc sách! Barnard cảm thấy rất lo âu và bối rối, bởi vì Washkansky rõ ràng không ý thức được rằng đây là một sự kiện quan trọng trong cuộc đời của chính ông mà còn là một sự kiện lịch sử trong y học. Nhưng bà vợ của Washkansky hỏi: "Cơ may mà bác sĩ nói là bao nhiêu?" Barnard trả lời: "80 phần trăm". Mười tám ngày sau cuộc giải phẫu, Washkansky qua đời. Ở đây, con số 80% được hiểu như thế nào? Có phải trong 100 người được phẫu thuật thì có 80 người sống sót? Hay nó phản ánh một độ tin tưởng, một cảm nhận cá nhân của bác sĩ Barnard? Phần dưới đây sẽ bàn đến ý nghĩa của con số nguy cơ này.

Khái niệm nguy cơ và xác suất

Tất cả những chỉ số vừa đề cập (RR, HR, OR) đều có liên hệ đến khái niệm *risk* (nguy cơ). Trong y khoa, *nguy cơ* là một khái niệm tương đối, nói đến *xác suất* của một sự kiện có thể xảy ra trong một thời gian nhất định. Do đó, có hai khía cạnh trong định nghĩa trên: sự kiện và thời gian. Sự kiện trong bối cảnh nghiên cứu y khoa đề cập đến những biến cố lâm sàng như tử vong, mắc bệnh, bệnh tái phát, v.v… Bởi vì nguy cơ có yếu tố thời gian, cho nên thông thường chúng ta nói đến nguy cơ là phải kèm theo nguy cơ trong một quãng thời gian nhất định. Nếu không có yếu tố thời gian, người ta có thể hiểu đó là nguy cơ trọn đời (lifetime risk).

Nguy cơ là một xác suất. Xác suất, theo Từ điển Tiếng Việt, là "*Số đo phần chắc của một biến cố ngẫu nhiên*." Định nghĩa này có ba phần: định lượng (*số đo*), chắc chắn, và ngẫu nhiên. Có gì mâu thuẫn ở đây: Chúng ta có thể nói một hiện tượng xảy ra do hai yếu tố, ngẫu nhiên và chắc chắn, chứ khó mà nói trong ngẫu nhiên không có sự chắc chắn! Vì thế chúng ta không thể dùng định nghĩa phổ thông này trong khoa học được.

Trong toán học, xác suất thường được giải thích bằng một ví vụ về trò chơi súc sắc hay một đồng xu. Một đồng tiền (kim loại) có hai mặt, và hãy tạm gọi hai mặt đó là A và B. Nếu chúng ta gieo đồng tiền đó một lần, kết quả sẽ là: hoặc A xuất hiện, hoặc B xuất hiện. Nếu chúng ta gieo nhiều lần (hàng triệu lần chẳng hạn), chúng ta sẽ kì vọng mặt A sẽ xuất hiện khoảng 50 phần trăm. Theo một định nghĩa của toán học, xác suất mặt A xuất hiện là 0.5 hay 50%. Chúng ta cũng có thể nói xác suất B xuất hiện là 50%.

Nhưng đó là một ví dụ tương đối vô bổ (rất hay thấy trong sách giáo khoa thống kê) chẳng có giá trị ứng dụng gì trong thực tế. Chúng ta có thể đặt ví dụ xác suất đó trong môi trường xã hội và y tế một cách thực tế hơn. Ví dụ trong một cộng đồng cư dân gồm 1000 người, có 20 người bị bệnh ung thư. Ở đây số người có tiềm năng bị ung thư là 1000 người, và số người thực sự bị ung thư (sự kiện) là 20 người. Chúng ta có vài cách phát biểu về qui mô này:

- Cách thứ nhất là dùng con số phần trăm: "*Tỉ lệ bệnh ung thư trong cộng đồng là 2 phần trăm*" (lấy 20 chia cho 1000, rồi lấy kết quả nhân cho 100).

- Cách thứ hai là đơn giản lấy 20 chia cho 1000 và với kết quả 0.02, chúng ta cũng có thể phát biểu "*Xác suất bị bệnh ung thư trong cộng đồng là 0.02*". Chúng ta cũng có thể thể hiện phát biểu đó bằng một kí hiệu toán học: P(ung thư) = 0.02. (P là viết tắt của chữ probability, tức xác suất).

- Cách thứ ba là dùng *tỉ số khả dĩ* (likelihood ratio) bằng cách lấy số người bị ung thư chia cho số người không bị ung thư: 20 / 980 = 0.082 (tỉ số khả dĩ bị ung thư là 0.082). Tỉ số khả dĩ càng cao, xác suất bị ung thư càng cao. (Chú ý: nếu tỉ số này là 1, điều đó có nghĩa là xác suất bị ung thư là 0.5).

Như vậy, *xác suất là tỉ lệ tần số một sự kiện xảy ra*. Đó là một định nghĩa xác suất cổ điển theo trường phái tiếng Anh gọi là *frequentist probability* (tần số xác suất). Nhưng cách định nghĩa dựa vào tần số như thế, dù rất thông dụng trong sách giáo khoa về toán học và thống kê xác suất, nó có vấn đề rất lớn trong việc diễn dịch cho một cá nhân. Chẳng hạn như nếu tôi nói "*Xác suất mà bạn bị ung thư là 0.10*" thì điều đó có nghĩa là gì? Nó có nghĩa là trong 100 người như bạn, có 10 người bị ung thư. Nói cách khác, nó là một con số áp dụng cho một quần thể, chứ không phải cho một cá nhân. Ấy thế mà câu phát biểu đó dùng cho một cá nhân! Do đó, có người cho rằng một phát biểu như thế hoàn toàn vô nghĩa, bởi vì một cá nhân là chỉ 1 cá nhân, mà 1 cá nhân thì không có mẫu số.

Một định nghĩa thứ hai được đề xuất từ thế kỉ 17 là trường phái *xác suất chủ quan* (subjective probability). Theo trường phái này, xác suất là một diễn đạt cá nhân. Chúng ta sử dụng xác suất hàng ngày nhưng không để ý. Chúng ta vẫn thường nói "*Hôm nay chắc trời mưa quá*", hay "*Tôi thấy anh hình như bị cảm lạnh*". Đó là những cảm nhận cá nhân về một sự kiện, một tình trạng, nhưng là những cảm nhận không chắc chắn (tức bất định). Cách phát biểu như trên là một cách diễn đạt mối liên hệ của một cá nhân đối với một sự kiện, nó không phải là một đặc tính khách quan của sự kiện. Chính vì thế mà có người đề nghị chúng ta nên nói "*xác suất về sự kiện*" (probability for an event), chứ không nên nói "*xác suất sự kiện*" (probability of an event). Xác suất, theo trường phái này, là một *số đo về sự bất định* (degree of uncertainty), hay một *số đo về mức độ tin tưởng* (degree of belief). Quay trở lại câu nói "*Xác suất mà bạn bị ung thư là 0.10*", theo cách hiểu này, là một cảm nhận chủ quan của cá nhân người phát biểu đến bệnh nhân. Không có cách gì để chứng minh câu phát biểu đó đúng hay sai (ngoại trừ xác suất là 0 hay 1).

Odds và xác suất

Tất cả diễn giải trên chỉ để nói một điểm: nguy cơ là xác suất, nhưng xác suất có thể thiểu theo trường phái chủ quan hay trường phái tần số.

Xác suất khác với *odds*. Odds là một khái niệm đặc thù trong văn hóa đánh bạc, và chỉ có người Anh mới có thuật ngữ odds, không có ngôn ngữ nào trên thế giới có chữ odds!

Nếu trong số 100 bệnh nhân có 10 người mắc bệnh trong một thời gian theo dõi, thì *nguy cơ* mắc bệnh (kí hiệu p) là:

$$p = 10 / 100 = 0.10$$

hay 10%. Nhưng odds được định nghĩa là:

$$odds = \frac{p}{1-p}$$

và trong ví dụ trên, chúng ta có:

$$odds = \frac{0.10}{0.90} = 0.11$$

Có nghĩa là cứ 11 người không mắc bệnh thì có 1 người mắc bệnh. Đương nhiên, nếu nguy cơ $p = 0.5$ (hay 50%) thì odds = 1, hay nếu $p = 0.9$, thì odds = 9. Nói cách khác, giá trị của xác suất hay nguy cơ dao động trong khoảng 0 và 1, nhưng odds không có giới hạn về giá trị, có thể gần 0 mà cũng có thể vô hạn. Nhưng cả hai đều là số dương. Qua ví dụ và định nghĩa trên, chúng ta thấy *odds* không phải là xác suất, và không thể xem là nguy cơ.

Định nghĩa OR, RR và HR

Các chỉ số OR, RR và HR đều đo lường mức độ tương quan (magnitude of association) giữa một yếu tố nguy cơ (risk factor) và nguy cơ mắc bệnh (risk of disease). Nhưng ý nghĩa thật của chúng có khi khác nhau. Để hiểu rõ các định nghĩa này, chúng ta sẽ xem qua kết quả của ba nghiên cứu sau đây (tôi sẽ chú trọng đến RR và OR, vì HR gần như là RR và sẽ được giải thích trong một phần sau):

Nghiên cứu 1: **zoledronic acid và gãy xương.** Trong một nghiên cứu gần đây về ảnh hưởng của zoledronic acid, một loại thuốc chống loãng xương và ngừa gãy xương trong gia đình bisphosphonates, các nhà nghiên cứu tuyển chọn 7736 phụ nữ sau mãn kinh, tuổi từ 65 đến 89 (Black DM, et al. *N Engl J Med* 2007 May 3;356(18):1809-22). Họ ngẫu nhiên chia các đối tượng nghiên cứu thành 2 nhóm: nhóm 1 gồm 3875 bệnh nhân được điều trị với zoledronate, và nhóm 2 gồm 3861 bệnh nhân trong nhóm đối chứng không được tiêm zoledronate mà chỉ uống calcium và vitamin D (còn gọi là nhóm chứng). Sau 3 năm theo dõi, có 92 người (tỉ lệ 2.4%) trong nhóm zoledronate gãy xương, và 310 người (hay 8.0%) trong nhóm chứng bị gãy xương đốt sống (vertebral fracture):

Bảng 1. Tóm lược kết quả nghiên cứu theo từng nhóm đối tượng sau 3 năm nghiên cứu về hiệu quả của zoledronic acid.

	Nhóm Zoledronate	Nhóm chứng (Placebo)
Không gãy xương	3783	3551
Gãy xương đốt sống	92	310
Tổng số	3875	3861

Nghiên cứu 2: **Tử vong trên tàu Titanic**. Ngày 10/4/1912, tàu du lịch Titanic chở 1309 du khách gặp nạn. Trong tai nạn này có 809 người không may mắn tử vong. Số người tử vong và sống sót được phân chia theo hạng vé như sau:

Bảng 2. Tai nạn tàu Titanic và tử vong theo hạng hành khách

	Chết	**Sống**
I	123	200
II + III	416	300

Tổng số	809	500

Nguồn: http://lib.stat.cmu.edu/S/Harrell/data/descriptions/titanic3info.txt

Nghiên cứu 3: **ung thư phổi và hút thuốc lá**. Một công trình nghiên cứu bệnh chứng mang tính lịch sử, vì đây là công trình lần đầu tiên cho thấy người hút thuốc lá có nguy cơ mắc bệnh ung thư phổi. Nghiên cứu do Richard Doll và Bradford Hill thực hiện. Họ chọn 649 người mắc bệnh ung thư phổi, và 649 người không mắc bệnh (nhóm chứng). sau đó, họ tìm hiểu tiền sử hút thuốc lá. Kết quả của công trình lịch sử này có thể tóm lược như sau (R Doll and B Hill. BMJ 1950; ii:739-748):

Bảng 3. Tần số hút thuốc lá trong nhóm ung thư phổi và không ung thư phổi.

	Ung thư phổi	**Nhóm chứng**
Hút thuốc lá	647	622
Không hút thuốc lá	2	27
Tổng số	649	649

Nghiên cứu 1 là nghiên cứu can thiệp trong mô hình randomized controlled trial (tức lâm sàng đối chứng ngẫu nhiên), nhưng nghiên cứu 2 và 3 là những nghiên cứu quan sát (tức không can thiệp). Có thể xem nghiên cứu 2 là một nghiên cứu cắt ngang (cross-sectional study hay một survey), và nghiên cứu 3 là nghiên cứu bệnh chứng.

Trong nghiên cứu 1, chúng ta có thể ước tính nguy cơ gãy xương cho nhóm điều trị (p_1) và nhóm chứng (p_2) như sau:

$$p_1 = \frac{92}{3875} = 0.024$$

và

$$p_2 = \frac{310}{3861} = 0.080$$

Relative risk (RR, hay có khi còn được gọi là *risk ratio*) được định nghĩa là tỉ số của hai nguy cơ:

$$RR = \frac{p_1}{p_2} = \frac{0.024}{0.080} = 0.295 \qquad [1]$$

Nói cách khác, nguy cơ gãy xương trong nhóm bệnh nhân điều trị bằng zoledronic acid giảm ~71% so với nhóm chứng.

Nhưng định nghĩa odds thì khác. Vẫn lấy ví dụ trên, odds của nhóm điều trị là (O_1):

$$O_1 = \frac{0.024}{1-0.024} = 0.0243$$

Và odds của nhóm chứng:

$$O_2 = \frac{0.080}{1-0.080} = 0.0872$$

Odds ratio (OR) được định nghĩa là tỉ số của hai odds:

$$OR = \frac{O_1}{O_2} = \frac{0.0243}{0.0872} = 0.279 \quad [2]$$

Chúng ta thấy trong trường hợp này, RR và OR không khác nhau đáng kể. Nhưng ý nghĩa của RR thì dễ diễn giải hơn vì chúng ta biết đó là tỉ số của 2 nguy cơ, còn ý nghĩa của OR thì phức tạp hơn, vì đó là tỉ số của 2 odds và odds không phải là nguy cơ hay xác suất!

RR và OR cũng có thể ước tính cho nghiên cứu cắt ngang như trường hợp tai nạn tàu Titanic. Nhưng trong trường hợp này, tỉ lệ tử vong không phải là một "incidence" (tỉ lệ phát sinh) mà là "prevalence" (tỉ lệ hiện hành) vì không có yếu tố thời gian ở đây. Gọi tỉ lệ tử vong cho nhóm hành khách hạng I là P_1 và hạng II+III là P_2, chúng ta có:

$$P_1 = \frac{123}{123+200} = 0.381$$

và

$$P_2 = \frac{416}{416+300} = 0.581$$

Do đó, RR là:

$$RR = \frac{P_1}{P_2} = \frac{0.381}{0.581} = 0.655 \quad [3]$$

Bởi vì đây là một nghiên cứu cắt ngang, P_1 và P_2 thực chất là prevalence, cho nên trong giới dịch tễ học không ai gọi là *RR* (như tôi vừa viết); họ có một thuật ngữ khác thích hợp hơn: đó là *prevalence ratio* (PR). Tuy nhiên, ngày nay, thuật ngữ prevalence ratio rất ít được sử dụng, và thay vào đó, người ta sử dụng một thuật ngữ chung khác có tên là *risk ratio*, vẫn viết tắt là RR. Nói cách khác, RR (risk ratio) là tên gọi chung cho relative risk và prevalence ratio.

Bây giờ, chúng ta thử tính OR theo trình tự trên. Trước hết là odds tử vong của nhóm hành khách hạng I:

$$O_1 = \frac{123}{200} = 0.615$$

Và odds tử vong cho nhóm hành khách hạng II và III:

$$O_2 = \frac{416}{300} = 1.387$$

Do đó, odds ratio tử vong của nhóm hành khách hạng I so với nhóm hạng II và III là:

$$OR = \frac{O_1}{O_2} = \frac{0.615}{1.387} = 0.443 \qquad [4]$$

Con số RR = 0.655 trên cho chúng ta biết hành khách hạng I có nguy cơ tử vong trong tàu Titanic thấp hơn hành khách hạng II và III khoảng 35% (lấy 1 trừ cho 0.655). Nhưng OR là 0.443, và nếu có người "can đảm" diễn giải rằng nguy cơ tử vong của nhóm hành khách hạng I thấp hơn nhóm hạng II và III là 56% thì sẽ rất sai. Sai là vì với OR chúng ta không có thể nói về *nguy cơ* hay *risk*, do đơn vị của OR là *odds*.

Tại sao trong nghiên cứu 1, RR và OR rất gần nhau, nhưng trong nghiên cứu 2 RR và OR quá khác nhau? Nhìn kĩ công thức tính RR và OR, chúng ta dễ dàng thấy một mối "liên hệ hữu cơ" (nếu bạn nào thích đại số có thể làm một chứng minh), và mối liên hệ này cho chúng ta thấy **nếu tỉ lệ (nguy cơ) bệnh thấp (như dưới $p < 0.1$) hay rất thấp ($p < 0.01$) thì OR rất gần với RR, nhưng khi nguy cơ bệnh cao (như trên $p > 0.2$) thì OR lúc nào cũng cao hơn RR**. Nói cách khác, OR lúc nào cũng cao hơn RR và mức độ khác biệt càng lớn khi nguy cơ bệnh càng cao (như nghiên cứu 2, khi *P* trên 0.35).

Bảng sau đây trình bày 10 nghiên cứu (tưởng tượng) mà tất cả đều có RR bằng 3, nhưng OR khác nhau. Như có thể thấy qua bảng này, khi nguy cơ mắc bệnh càng cao thì OR càng cao hơn RR.

Bảng 4. So sánh giữa OR và RR theo tần số mắc bệnh

Nghiên cứu	Nguy cơ mắc bệnh		Odds of disease		So sánh giữa RR và OR	
	Nhóm 1 (p_1)	Nhóm 2 (p_2)	Nhóm 1 (O_1)	Nhóm 2 (O_2)	RR	OR
1	0.001	0.003	0.002	0.003	3	3.01
2	0.01	0.03	0.01	0.03	3	3.06
3	0.02	0.06	0.02	0.06	3	3.13
4	0.05	0.15	0.05	0.18	3	3.35
5	0.10	0.30	0.11	0.43	3	3.86

6	0.15	0.45	0.18	0.82	3	4.64
7	0.20	0.60	0.25	1.50	3	6.00
8	0.25	0.75	0.33	3.00	3	9.00
9	0.30	0.90	0.43	9.00	3	21.0
10	0.33	0.99	0.49	99.0	3	2101.0

Thế thì câu hỏi đặt ra là tại sao chúng ta cần OR? Tại vì có một số mô hình nghiên cứu mà chúng ta không thể ước tính RR, mà chỉ có thể ước tính OR. Một trong những mô hình nghiên cứu đó chính là nghiên cứu bệnh chứng (nghiên cứu 3). Trong nghiên cứu bệnh chứng, chúng ta biết trước có bao nhiêu người mắc bệnh và không mắc bệnh, và chúng ta đi ngược quá khứ để tìm hiểu yếu tố nguy cơ. Do đó, chúng ta không thể ước tính nguy cơ, nhưng có thể ước tính *odds*.

Quay lại với kết quả của nghiên cứu 3, chúng ta tính toán odds hút thuốc lá cho nhóm ung thư phổi (gọi tắt là O_1):

$$O_1 = \frac{647}{622} = 323.5$$

Và odds hút thuốc lá cho nhóm chứng:

$$O_2 = \frac{622}{22} = 23.04$$

Từ đó, OR được ước tính:

$$OR = \frac{O_1}{O_2} = \frac{323.5}{23.04} = 14.04$$

Chỉ số OR này được diễn giải như sau: odds ung thư phổi trong nhóm hút thuốc lá cao hơn nhóm không hút thuốc lá 14 lần. Xin nhấn mạnh rằng chúng ta chỉ nói odds chứ không phải risk (nguy cơ).

Nhưng như chúng ta vừa thấy qua nghiên cứu 1, khi tần số bệnh trong cộng đồng thấp thì OR có thể xem là RR. Chúng ta biết rằng ung thư phổi là một bệnh có tần số thấp trong cộng đồng, cho nên OR = 14 *có thể* diễn giải rằng *nguy cơ* mắc bệnh ung thư trong nhóm hút thuốc lá cao hơn nhóm không hút thuốc lá 14 lần.

Khái niệm về hazard

Hazard thường được dịch sang tiếng Việt là "may rủi" hay "rủi ro". Nhưng trong thuật ngữ dịch tễ học có nghĩa là *lực nguy cơ*, vì nhà dân số học Gompertz sử dụng từ này lần đầu vào năm 1825 để chỉ cái mà ông gọi là "Force of mortality" (lực tử vong). Ngày nay, ít ai sử dụng force of mortality mà chỉ nói "nguy cơ". Do đó, hazard ratio (HR)

cũng có khi được xem là risk ratio (RR). Có thể xem *incidence* là tương đương với khái niệm *tốc độ* (velocity) bên vật lí, và *hazard* là gia tốc (acceleration) của vật lí.

Để hiểu hazard ratio, chúng ta thử xem qua một nghiên cứu dân số dịch tễ học mà trong đó các nhà nghiên cứu ước tính tỉ lệ tử vong của những người trong độ tuổi 50-54, 55-59, 60-64 và 65-69 như sau:

Bảng 5. Tỉ lệ tử vong tính trên 100,000 dân số Mĩ 1990

Nguyên nhân tử vong	50-54	55-59	60-64	65-69
Ung thư phổi	91.1	176.0	289.9	399.1
Tim mạch	276.1	473.6	766.6	1197.2

Câu hỏi đặt ra là: *đối với những đàn ông 50 tuổi, nguy cơ tử vong trong vòng 10 năm vì ung thư phổi so với bệnh tim mạch là bao nhiêu*? Nhưng trong nghiên cứu trên, chúng ta chỉ theo dõi tử vong trong vòng 1 năm, cho nên câu hỏi trên khó có câu trả lời chính xác. Tuy nhiên, điều may mắn là các nhà nghiên cứu có nhiều độ tuổi khác nhau. Do đó, chúng ta có thể ước tính nguy cơ tích lũy (cumulative risk). Hàm số nguy cơ tích lũy có thể hơi phức tạp một chút nhưng có thể viết đơn giản như sau: $U_k(T) = \int_0^T \lambda_k(t)dt$; trong đó, hàm $\lambda_k(t)$ là hàm tử vong, T = 10 năm (theo câu hỏi tính từ 50 tuổi). Qua vài thao tác đại số, chúng ta có thể ước tính xác suất tử vong cho người 50 tuổi trong vòng 10 năm vì bệnh ung thư phổi (kí hiệu P_1) như sau:

$$P_1 = 1 - \exp\left[-\frac{(91.1\times 5)+(176.0\times 5)}{100{,}000}\right] = 0.013266$$

Và nguy cơ tử vong vì bệnh tim mạch là:

$$P_2 = 1 - \exp\left[-\frac{(276.1\times 5)+(473.6\times 5)}{100{,}000}\right] = 0.0376791$$

Chú ý trong cách tính trên, chúng ta phải nhân tỉ lệ tử vong cho 5 vì trong mỗi nhóm tuổi, khoảng cách là 5 (như 50 đến 54). Mẫu số 100,000 là vì tỉ lệ trong bảng trên được tính trên 100,000 dân số.

Từ đó, hazard ratio được định nghĩa là tỉ số của hai nguy cơ tích lũy:

$$HR = \frac{P_1}{P_2} = \frac{0.013266}{0.037679} = 0.36$$

Nói cách khác, nguy cơ tử vong vì ung thư phổi thấp hơn nguy cơ tử vong vì bệnh tim khoảng 64%.

Tóm lược

Để kết thúc bài này, Bảng 6 sau đây tóm lược mối liên hệ giữa OR, RR và HR. Tôi muốn nhắc lại một số điểm chính như sau:

Khái niệm nguy cơ là xác suất, khác với khái niệm odds không phải là xác suất mà chỉ là một tỉ số. Do đó, về diễn giải, OR không thể diễn giải bằng ngôn ngữ nguy cơ, nhưng trong vài trường hợp thì OR cũng *có thể* xem là RR.

Trong dịch tễ học, chúng ta muốn ước tính *relative risk* (RR), còn có khi gọi là *risk ratio.* Nhưng vì trong nghiên cứu bệnh chứng, chúng ta không thể ước tính RR, nên phải ước tính OR. Do đó, OR chính là một ước số (estimate) của RR.

Nhưng vì OR rất dễ tính, nên **OR có thể tính cho bất cứ mô hình nghiên cứu nào** (như nghiên cứu bệnh chứng, cắt ngang, hay nghiên cứu xuôi theo thời gian -- prospective study). Nhưng với các nghiên cứu mà tần số bệnh khá cao (trên 10%) thì OR lúc nào cũng ước tính RR cao hơn thực tế (over-estimation). Chính vì thế mà ngày nay, những ai hiểu vấn đề không sử dụng OR hay hồi qui logistic cho các nghiên cứu như thế. Tuy nhiên, để tính prevalence ratio hay risk ratio trong các nghiên cứu như thế đòi hỏi phải biết sử dụng chương trình R để phân tích thích hợp.

RR (relative risk hay risk ratio) chỉ có thể ước tính cho nghiên cứu xuôi theo thời gian hay cắt ngang, chứ không thể tính từ nghiên cứu bệnh chứng. HR chỉ có thể ước tính cho các nghiên cứu xuôi theo thời gian với điều kiện chúng ta biết chính xác thời điểm hay thời gian mà biến cố lâm sàng xảy ra.

Về mặt thống kê, OR được ước tính qua mô hình hồi qui logistic; RR ước tính qua mô hình hồi qui Poisson hay hồi qui nhị phân; và HR thì ước tính chỉ qua mô hình hồi qui Cox (Cox's proportional hazard model). Tôi sẽ bàn qua cách phân tích và diễn giải các mô hình nhị phân trong một bài sau, nhưng phần phụ chú dưới đây giải thích sơ qua về mô hình Poisson.

Bảng 6. Phân biệt một số chỉ số đo lường ảnh hưởng, thuật ngữ và phương pháp phân tích

Chỉ số	Nguồn gốc	Tên gọi khác	Luật phân phối	Phương pháp phân tích
Odds ratio	Odds ratio (Cornfield, 1951)	Cross-product ratio	Nhị phân (binomial distribution)	Hồi qui logistic (Cox, 1958)
Relative risk	Relative risk	Risk ratio	Poisson hay	Poisson

	(Cornfield, 1951)		Nhị phân	regression (Cochran, 1940; Kutner và Beauchamp 1973)
Hazard ratio	Intensity of mortality ratio (Gompertz 1825)	Force of mortality	Exponential, Poisson	Cox's proportional hazards (Cox 1972)

Phụ chú:

Một dạng nghiên cứu khác cũng có thể sử dụng risk ratio qua ví dụ sau đây. Số liệu sau đây là số trường hợp mắc bệnh ung thư da ở phụ nữ thuộc "thành phố song sinh" Minneapolis – St Paul (Mĩ):

Độ tuổi	Số trường hợp ung thư da	Dân số	Tỉ lệ phát sinh trên 1000 dân số
15-24	1	172,675	0.0058
25-34	16	146,207	0.1094
35-44	30	121,374	0.2472
45-54	71	111,353	0.6376
55-64	102	83,004	1.2289
65-74	130	55,932	2.3243
75-84	133	29,007	4.5851
85+	40	7,538	5.3064

Dựa vào hai số liệu trên, chúng ta có thể ước tính tỉ lệ phát sinh (incidence rate) tính trên 1000 dân số, và tỉ lệ này được trình bày trong cột cuối cùng của bảng số liệu. Chúng ta dễ dàng thấy tỉ lệ phát sinh ung thư da tăng theo độ tuổi.

Chúng ta cần một mô hình để mô tả xu hướng trên, hay mối liên hệ giữa độ tuổi và tỉ lệ phát sinh. Ở đây, có một khó khăn là dân số khác nhau giữa các độ tuổi, mà tỉ lệ thì tùy thuộc vào độ tuổi. Cho nên chúng ta cần một mô hình có thể điều chỉnh cho hai yếu tố này. Một mô hình thích hợp cho trường hợp này là mô hình hồi qui Poisson.

Gọi N_i là dân số của độ tuổi i^{th} (cột số 3 trong bảng trên), và gọi μ_i là số trường hợp ung thư da (cột số 2), chúng ta có thể xem tỉ số $\frac{\mu_i}{N_i}$ (cột số 4) như là một tỉ lệ phát sinh bệnh. Bây giờ, để tiện cho việc tính toán, chúng ta hoán chuyển tỉ số này sang đơn vị logarít:

$$\log\left(\frac{\mu_i}{N_i}\right) = \log(\mu_i) - \log(N_i) \qquad [1]$$

Mô hình này phát biểu rằng log của tỉ lệ phát sinh, tức là $\log\left(\frac{\mu_i}{N_i}\right)$, là một hàm số của độ tuổi (và chúng ta kí hiệu độ tuổi bằng x). Nói cách khác, mô hình này phát biểu rằng:

$$\log\left(\frac{\mu_i}{N_i}\right) = \alpha + \beta x_i \qquad [2]$$

Thay thế [1] vào vế trái của [2], chúng ta có:

$$\log(\mu_i) - \log(N_i) = \alpha + \beta x_i$$

hay nói cách khác:

$$\log(\mu_i) = \alpha + \beta x_i + \log(N_i) \qquad [3]$$

Mô hình [3] chính là mô hình hồi qui Poisson (Poisson Regression). Cách diễn đạt mô hình theo công thức [3] cho phép chúng ta xây dựng một hàm số khả dĩ (likelihood function) để ước tính các thông số α và β. Trong mô hình trên, $\log(N_i)$ được gọi là *offset,* tức là "điểm nhấn".

Chúng ta có thể sử dụng R để ước tính các thông số trong mô hình [3] như sau:

```
age = c(19.5,29.5,39.5,49.5,59.5,69.5,79.5, 89.4)
cases = c(1, 16, 30, 71, 102, 130, 133, 40)
pop = c(172675, 123065, 96216, 92051, 72159, 54722, 32185, 8328)
dataset = data.frame(age, cases, pop)
fit = glm(cases ~ age + offset(log(pop)), family=poisson,data=dataset)
summary(fit)
```

Kết quả của phân tích là:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.87198  -1.67519  -0.07185   1.20816   1.99291

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -10.551604   0.168780  -62.52   <2e-16 ***
age           0.063629   0.002475   25.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 845.789  on 7  degrees of freedom
Residual deviance:  44.288  on 6  degrees of freedom
AIC: 91.688

Number of Fisher Scoring iterations: 5
```

Chúng ta chỉ chú ý đến thông số `intercept` và `age`. Thay thế các thông số này vào mô hình [3] chúng ta có:

$$\log(\mu_i) = -10.552 + 0.064 x_i + \log(N_i)$$

trong đó, *x* là độ tuổi.

Cũng có thể viết mô hình trên thành tỉ lệ phát sinh (thay vì log của tỉ lệ) như sau:

$$\frac{\mu_i}{N_i} = e^{-10.552+0.064x}$$

Nếu tuổi là 19.5 ($x = 19.5$), chúng ta có thể ước tính tỉ lệ phát sinh là:

$$\frac{\mu_i}{N_i} = e^{-10.552+0.064\times 19.5} = 0.0000904$$

Nếu tuổi là 29.5, tỉ lệ phát sinh là:

$$\frac{\mu_i}{N_i} = e^{-10.552+0.064\times 29.5} = 0.000171$$

Tỉ số của hai tỉ lệ phát sinh này là: 0.000171 / 0.0000904 = 1.89. Nói cách khác, tỉ lệ phát sinh tăng gấp 1.89 lần cho mỗi 10 tuổi tăng. Con số 1.89 này còn có tên là *risk ratio.*

Thật ra, chúng ta có thể ước tính ngay từ ước số của R:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.551604   0.168780  -62.52   <2e-16 ***
age           0.063629   0.002475   25.71   <2e-16 ***
```

Trong phần trên, $\beta = 0.0636$ có nghĩa là cứ mỗi một tuổi tăng thì tỉ số nguy cơ tương đối tăng $e^{0.0636} = 1.065$ (hay tăng 6.5%). Do đó, khi tăng 10 tuổi thì nguy cơ tương đối mắc bệnh tăng $e^{0.0636 \times 10} = 1.89$, hay 89%.