Lâm sàng thống kê

Không nên nhầm lẫn giữa P(R | D) và P(D | R)!

Nguyễn Văn Tuấn

Vấn đề

Trong vụ bệnh tả bộc phát vừa qua ở các tỉnh phía Bắc, có một phát biểu làm "ám ảnh" tôi khá lâu: đó là phát biểu của các quan chức y tế (cụ thể là giám đốc Sở Y tế Hà Nội) về mối liên hệ giữa mắm tôm và bệnh tả. Ngày 30/10, giám đốc Sở Y tế Hà Nội cho biết có 30 trường hợp đã nhập viện để điều trị và "*Nguyên nhân chính dẫn đến căn bệnh này là do người bệnh ăn thực phẩm sống như: mắm tôm, mắm tép, tiết canh, gỏi hải sản… trong đó có tới 90% số người mắc bệnh là do ăn mắm tôm sống*." Liền theo đó, trong ngày 30/10 UBND Thành phố Hà nội ra công lệnh tạm thời cấm sản xuất, vận chuyển, buôn bán và tiêu thụ mắm tôm trên địa bàn thành phố dù "*Trong khi chưa xác định được nguyên nhân chính gây ra hàng loạt ca tiêu chảy cấp*".

Trả lời chất vấn của một đại biểu trong Quốc hội, Bộ trưởng Bộ Y tế cũng nói rằng "Hội đồng chuyên môn" căn cứ vào 93% người bệnh ăn mắm tôm cũng như "tiền sử" thứ nước chấm này đã từng gây ra dịch những lần trước nên ra quyết định cấm mắm tôm.

Đây là một vấn đề lí thú liên quan đến triết lí khoa học và xác suất thống kê học. Triết lí khoa học cung cấp cho chúng ta những cơ sở lí luận để phân biệt giữa *nguyên nhân* (cause), *yếu tố nguy cơ* (risk factor), và *hệ quả* (outcome). Môn xác suất cung cấp cho chúng ta phương tiện toán học để diễn dịch các vấn đề triết lí thành các vấn đề định lượng để chúng ta hiểu được. Bài này không bàn về triết lí (vì "*không muốn dùng những danh từ hoa mĩ / để chỉ nói suông những triết lí cuộc đời*" – xin trích thơ của Ma Văn Kháng), mà chỉ quan tâm đến vấn đề xác suất định lượng trong câu phát biểu trên, và bàn ngắn gọn về ý nghĩa của câu nói đó trong thực tế lâm sàng cũng như y học cộng đồng.

Kí hiệu

Bàn chuyện khoa học cần đến kí hiệu, không phải để làm dáng trí thức hay để làm cho người ta khó hiểu, mà để khỏi phải lặp đi lặp lại câu chữ dài dòng. Chúng ta tạm đặt vài kí hiệu:

- Gọi *bệnh tả* là hệ quả và cho kí hiệu *D* (viết tắt chữ *disease*);
- Tương tự, gọi *ND* (non-disease) là tình trạng không mắc bệnh;
- Gọi mắm tôm là yếu tố nguy cơ, và kí hiệu bằng *R* (viết tắt chữ risk factor);
- Gọi *NR* (non-risk factor) mô tả không ăn mắm tôm;

- Và sau cùng, gọi *P* là xác suất của một sự kiện.

Xin nhắc lại khái niệm căn bản: trong y học, xác suất của một sự kiện là tỉ lệ mà sự kiện đó xảy ra trong một thời gian và không gian xác định. Chẳng hạn như khi chúng ta viết "P(D) = 0.2" thì biểu thức này có nghĩa là xác suất bệnh tả xảy ra là 20%, hay hiểu theo trường phái tần số (frequentism) thì cứ 100 người có 20 người mắc bệnh.

Xác suất có điều kiện và không điều kiện

Có hai loại xác suất: *xác suất có điều kiện* (conditional probability) và *xác suất vô điều kiện* (unconditional probability). Xác suất có điều kiện thường kèm theo chữ "nếu" hay dấu "|" (tiếng Anh đọc dấu này là "given" hay "if"). Chẳng hạn như *P(D)* là xác suất không có điều kiện, vì nó đề cập đến tỉ lệ mắc bệnh trong cộng đồng, hay *P(R)* cũng là xác suất không có điều kiện, vì nó cho biết trong cộng đồng có bao nhiêu người ăn mắm tôm mà không nói đến một nhóm cá biệt nào cả.

Khi ông Bộ trưởng nói 93% người mắc bệnh tả từng ăn mắm tôm là một phát biểu xác suất có điều kiện. Theo kí hiệu xác suất và theo định nghĩa tôi trình bày trên, câu ông nói tương đương với biểu thức:

$$P(R \mid D) = 0.93 \qquad [1]$$

Nhưng phát biểu "trong số 2000 bệnh nhân, có 1860 hay 93% từng ăn mắm tôm" là một sự thật, một quan sát. Một sự thật có ý nghĩa cần phải đặt nó trong bối cảnh. Nói "ăn mắm tôm" mà không nói đến liều lượng ăn, thời gian quen ăn, và các thực phẩm khác cùng đi với mắm tôm thì câu nói đó ("ăn mắm tôm") hoàn toàn không có nghĩa gì đáng chú ý.

Phát biểu "trong số 2000 bệnh nhân, có 1860 hay 93% từng ăn mắm tôm" mang tính tuy tìm về quá khứ. Anh *đang* mắc bệnh, tôi hỏi anh *đã* ăn gì. Nhưng một câu hỏi như thế, dù giúp ích cho chẩn đoán và đôi khi liệu pháp điều trị, nhưng mang tính "hàn lâm" hơn là thực tế. Bệnh nhân có thể nói: tôi *đang* mắc bệnh, anh làm ơn điều trị cho tôi. Nói cách khác, tôi đang nằm trên giường bệnh và chờ điều trị, còn quá khứ tôi ăn gì tuy cũng quan trọng nhưng chưa cần thiết ngay lúc này.

Điều cần thiết ngay lúc này là những người chưa mắc bệnh, nhưng có ăn mắm tôm. Câu hỏi của họ là: chúng tôi cũng từng ăn mắm tôm, vậy nguy cơ mắc bệnh của tôi là bao nhiêu. Nói theo ngôn ngữ xác suất và theo kí hiệu định nghĩa trên, câu hỏi này là:

$$P\left(D \mid R\right) = ? \qquad [2]$$

Chú ý kĩ: *P*(*D* | *R*) *rất* khác với *P*(*R* | *D*). Xác suất, nói cho cùng, là một phân số. Phân số thì phải có tử số và mẫu số. Phát biểu của ông Bộ trưởng, *P(R | D),* mẫu số là *D*, tức là những người mắc bệnh; và tử số là *R*, tức những người ăn mắm tôm. Nhưng

với $P(D \mid R)$, số người ăn mắm tôm là mẫu số và số người mắc bệnh là tử số. Có thể thể hiện phân biệt giữa hai biểu thức này qua một bản thống kê dưới đây:

Bảng số liệu để phân biệt P(D | R) và P(R | D)

	Mắc bệnh tả (D)	Không mắc bệnh tả (ND)	Tổng số
Ăn mắm tôm (R)	*a*	*b*	*mt*
Không ăn mắm tôm (NR)	*c*	*d*	*kmt*
Tổng số	*bt*	*kbt*	*N*

Ghi chú: Trong bảng trên, ***a*** là số người mắc bệnh và có ăn mắm tôm; ***b*** là số người không mắc bệnh và có ăn mắm tôm; ***c*** là số người mắc bệnh và không ăm mắm tôm; và ***d*** là số người không mắc bệnh và cũng không ăn mắm tôm. Ngoài ra, ***mt*** là tổng số người ăn mắm tôm trong cộng đồng; ***kmt*** là tổng số người không ăn mắm tôm trong cộng đồng; ***bt*** là số người mắc bệnh tả; ***kbt*** là số người không mắc bệnh tả; và ***N*** là tổng số dân trong cộng đồng.

Do đó, theo bảng trên, chúng ta có:

$$P(D \mid R) = \frac{a}{bt}$$

Và:

$$P(R \mid D) = \frac{a}{mt}$$

Trong dịch tễ học và chẩn đoán lâm sàng, $P(D \mid R) = \frac{a}{bt}$ được gọi là độ nhậy (sensitivity), còn $P(R \mid D) = \frac{a}{mt}$ là [hơi dài dòng chút] *trị số tiên lượng dương tính* (positive predictive value).

Giải đáp vấn đề

Bảng số liệu trên hàm ý nói chúng ta phải có số liệu *a, b, c,* và *d* cho toàn bộ dân số phía Bắc. Trong thực tế, chúng ta không / chưa có điều kiện để làm thống kê như thế. Tuy nhiên, chúng ta cần phải sử dụng nghiệp vụ của giới công an để truy tìm dữ liệu!

Chúng ta biết rằng số người mắc bệnh tiêu chảy khoảng 2000 người (hi vọng là tất cả những người mắc bệnh đã khai báo với địa phương, nhưng ở đây chúng ta đang bàn chuyện lí thuyết nên điều này cũng chưa quan trọng). Bộ Y tế cũng cho chúng ta biết rằng trong số này có khoảng 15% nhiễm vi khuẩn tả. Do đó, số người mắc bệnh tả phải vào khoảng:

$$bt = 2000 \times 0.15 = 300$$

Chúng ta cũng biết qua câu phát biểu của Bộ trưởng rằng 93% người mắc bệnh từng ăn mắm tôm, cho nên *a* có thể ước tính như sau:

$$a = 300 \times 0.93 = 279$$

Suy ra, số người mắc bệnh tả nhưng không ăn mắm tôm là $c = 300 - 279 = 21$ người. Tóm lại:

Bảng số liệu để phân biệt P(D | R) và P(R | D)			
	Mắc bệnh tả (D)	Không mắc bệnh tả (ND)	Tổng số
Ăn mắm tôm (R)	279	b	mt
Không ăn mắm tôm (NR)	21	d	kmt
Tổng số	300	kbt	N

Chúng ta đã "giải" được cột thứ nhất trong bảng số liệu trên. Bây giờ cột thứ hai có thể suy luận từ cột thứ 3 (tổng số). Chúng ta không biết bao nhiêu người không mắc bệnh, nhưng có thể (chỉ "có thể" thôi) biết bao người trong cộng đồng phía Bắc bằng cách dựa vào số liệu dân số. Ở đây, tôi không có con số đó, nhưng vì lí do minh họa, cứ giả định là $N = 1.000.000$ người. Như vậy số người không mắc bệnh tả phải là 1,000,000 – 300 = 999,700.

Đến đây thì có thêm vấn đề: chúng ta không biết trong số không mắc bệnh này có bao nhiêu người ăn mắm tôm. Nhưng cũng lại vì lí do minh họa, chúng ta có thể suy luận rằng phần lớn người Bắc ăn thịt chó, mà ăn thịt chó thì phải có mắm tôm. Hai chữ "phần lớn" ở đây có thể định lượng là 80%. Vậy thì trong số 1,000,000 người, ắt phải có khoảng *mt* = 800,000 người ăn mắm tôm.

Nhưng trong số 800,000 người không ăn mắm tôm, chúng ta đã biết có 279 người mắc bệnh. Vậy thì số người ăn mắm tôm mà không mắc bệnh phải là $b = 800,000 - 279 = 799,721$. Bây giờ chúng ta "cập nhật hóa" bản g số liệu trên:

Bảng số liệu để phân biệt P(D | R) và P(R | D)			
	Mắc bệnh tả (D)	Không mắc bệnh tả (ND)	Tổng số
Ăn mắm tôm (R)	279	799,721	800,000
Không ăn mắm tôm (NR)	21	d	kmt
Tổng số	300	999,700	1,000,000

Từ đó, dễ suy ra rằng, số người không ăn mắm tôm và không mắc bệnh là d = 1,000,000 – 800,000 – 21 = 199,979. Bây giờ thì chúng ta có một bảng số liệu hoàn chỉnh:

Bảng số liệu để phân biệt P(D | R) và P(R | D)

	Mắc bệnh tả (D)	Không mắc bệnh tả (ND)	Tổng số
Ăn mắm tôm (R)	279	799,721	800,000
Không ăn mắm tôm (NR)	21	199,979	200,000
Tổng số	300	999,700	1,000,000

Bây giờ với các số liệu trên, chúng ta có thể ước tính xác suất (tỉ lệ) những người ăn mắm tôm và mắc bệnh tả:

$$P(D \mid R) = \frac{279}{800,000} = 0.000349$$

Một nguy cơ rất thấp! Nói cách khác, cứ 100,000 người có khoảng 35 người mắc bệnh. Nhưng chúng ta cần phải so sánh với xác suất những người không ăn mắm tôm mà cũng mắc bệnh, và xác suất này là:

$$P(D \mid NR) = \frac{21}{200,000} = 0.000105$$

Do đó, nguy cơ mắc bệnh ở những người ăn mắm tôm so với những người không ăn mắm tôm là:

$$RR = \frac{0.000349}{0.000105} = 3.32$$

Đó chính là tỉ số nguy cơ (hay còn gọi là *relative risk*, RR). Một tỉ số nguy cơ 3.32 cũng đáng quan tâm, nhưng chưa đủ để kết tội mắm tôm là nguyên nhân. Tuy nhiên, các tính toán trên đây chỉ để minh họa cho sự khác biệt giữa *P(D | R)* và *P(R | D)*, chứ không có mục đích tranh luận nguyên nhân và yếu tố nguy cơ ở đây.

Quay lại câu phát biểu của Bộ trưởng

Quay lại câu phát biểu của Bộ trưởng Bộ Y tế: "*Cho đến giờ, căn cứ lâm sàng đợt đầu tiên 93% là ăn mắm tôm, không có gì khác. […] Do vậy, hội đồng chuyên môn đã kiến nghị trong thời gian dịch không dùng mắm tôm sống*." Câu này có nghĩa gì? Câu

phát biểu không đầy đủ và hơi mất chuẩn tiếng Việt! Câu "93% là ăn mắm tôm" không có nghĩa gì cả. 93% của quần thể nào?

Tuy nhiên, chúng ta hiểu ông muốn nói rằng trong tổng số bệnh tả 93% người từng ăn mắm tôm. Nói cách khác, $P(R \mid D) = 0.93$. Câu này không cho chúng ta biết gì về bệnh, mà chỉ cho chúng ta biết về "tiền sử" của mắm tôm ở những người bệnh tả, do mẫu tự *R* xuất hiện trước mẫu tự *D*.

Cố nhiên câu phát biểu trên khác với câu "trong số những người ăn mắm tôm, có 93% mắc bệnh", tức $P(D \mid R) = 0.93$. Câu này quan trọng và có ý nghĩa hơn, vì nó rõ ràng cho thấy mắm tôm nguy hiểm.

Nhưng chúng ta vẫn cần phải biết tỉ số nguy cơ để biết mức độ (magnitude) của mối liên hệ giữa mắm tôm và bệnh tả. Chúng ta cần số liệu nghiên cứu nhiều hơn nữa để có thể tiến đến một chính sách tối ưu.

Vấn đề là chúng ta không có điều kiện và tiền bạc để làm một nghiên cứu trên hàng triệu người như bảng số liệu minh họa trên giả định. Nhưng chúng ta có thể tiến hành một nghiên cứu ít tốn kém hơn, dễ dàng hơn, mà vẫn có thể ước tính tỉ số nguy cơ: đó chính là nghiên cứu bệnh chứng (case-control study). Nhưng thiết kế một nghiên cứu bệnh chứng không đơn giản, nhất là trong bối cảnh có rất nhiều yếu tố nguy cơ có thể dẫn đến bệnh tả. Nhưng đây là vấn đề phương pháp không nằm trong phạm vi bài lâm sàng thống kê này, nên tôi xin bàn vào một dịp khác.

Nói tóm lại, cần phân biệt $P(R \mid D)$ và $P(D \mid R)$! Không phân biệt được hai phát biểu này có thể dẫn đến sai lầm nghiêm trọng [1] và chúng ta đã thấy sai lầm qua cách "kết tội" mắm tôm của các đồng nghiệp (nói đúng hơn là quan chức) y tế.

Chú thích:

[1] Câu phát biểu "93% người mắc bệnh có tiền sử ăn mắm tôm, suy ra mắm tôm là thủ phạm gây bệnh" (xin lỗi các bạn tôi lặp lại câu này một lần nữa) còn được gọi là "Prosecutor Fallacy", tức là ngụy biện của công tố viên. Kiểu lí luận này đã gây ra khá nhiều tai hại cho nhiều nạn nhân, và trường hợp sau đây là một ví dụ tiêu biểu:

Sally Clark là một nữ luật sư, thuộc thành phần trung lưu ở Anh. Bà hạ sinh hai người con nhưng cả hai đều chết sau khi sinh: đứa con đầu chết lúc 11 tuần tuổi (1996), đứa con thứ hai chết lúc 8 tuần sau khi sinh (1997). Năm 1998, bà bị cảnh sát tố cáo là thủ phạm giết con và đưa bà ra tòa.

Trong phiên tòa xét xử, Tòa mời Ngài (sir) Roy Meadow, một giáo sư nhi khoa rất nổi tiếng, làm nhân chứng chuyên gia (expert witness). Giáo sư Meadow lí giải rằng nghiên cứu cho thấy xác suất một trẻ sơ sinh chết do đột tử (SIDS hay sudden infant death syndrome) khoảng 1 trên 8500. Ông lí giải tiếp rằng nếu 2 trẻ cùng chết trong một gia đình thì xác suất là 1/8500 x 1/8500 và kết quả là khoảng 1 trên 73 triệu. Nói cách khác, ông cho rằng xác suất mà Sally Clark có tội là 1 trừ cho 1/73.000.000 = 0.9999999…9 (tức bằng 1 hay 100%). Tòa án tuyên án Sally Clark có tội và đi tù.

Khi tòa án hỏi có cần chuyên gia thống kê làm chứng hay không thì cả công tố viên và giáo sư đều nói rằng đây không phải là "rocket science" (khoa học không gian) nên không cần các chuyên gia thống kê. Họ tự tin rằng họ rành về thống kê và xác suất. Toà án dự vào con số của Meadow và tuyên án chung thân cho bà Sally Clark.

Nhưng khi sự việc được báo chí tường thuật, các nhà thống kê học bắt đầu chú ý, và họ tìm hiểu kĩ hơn. Họ phát hiện rằng giáo sư Meadow phạm một số sai lầm cực kì cơ bản của xác suất. Sai lầm thứ nhất là ông giả định rằng xác suất 2 trẻ em chết trong một gia đình độc lập với nhau (nên nhân 2 xác suất với nhau). Điều này không đúng, vì đột tử có thể có nguyên nhân từ môi trường và di truyền, mà hai em là anh em, tức có thể có cùng gien và cùng mẹ (cùng môi trường) nên 2 hiện tượng không thể độc lập. Thật ra, nếu 1 trẻ đã chết vì độ tử thì xác suất trẻ thứ 2 chết vì độ tử rất cao. Do đó, nhân 2 xác suất trong trường hợp này hoàn toàn sai lầm. Sai lầm thứ hai là giáo sư Meadow không định được xác suất một phụ nữ như bà Sally Clark có thể phạm tội giết con. Đáng lẽ họ phải tính xác suất mà một người trung lưu như bà Clark giết con là bao nhiêu, tức họ cần tính P(giết con | trung lưu), chứ không nên tính P(trung lưu | giết con).

Ngày 29/1/2003, sau khi kháng án với nhân chứng từ một giáo sư thống kê học, tòa tuyên án Sally Clark không có tội. Tòa cũng kỉ luật giáo sư Meadow vì đưa bằng chứng sai. Giáo sư Meadow sau đó bị tước chức danh và cấm hành nghề y khoa. Bà Clark vừa qua đời vào tháng 3 năm 2007, thọ 42 tuổi.

Bạn đọc có thể tìm hiểu về vụ này qua trang web http://www.sallyclark.org.uk. Ngoài ra, còn có rất nhiều bài báo khoa học trên các tập san y khoa và khoa học như *Lancet, British Medical Journal, Science, Nature*, *Statistics*, *Epidemiology*, *Medical Journal of Australia*, v.v… bàn về trường hợp này. Bài trên MJA tóm lược khá đầy đủ.

Năm nay, một trường hợp khác ở Hà Lan liên quan đến y tá Lucia de Berk cũng do sai lầm về diễn dịch xác suất có điều kiện mà tôi bàn trong bài này. De Berk bị tố cáo giết chết 4 bệnh nhân trong khi trực ở bệnh viện, và bị tòa án tuyên án chung thân vào năm 2003. Bằng chứng kết tội de Berk gần như dựa vào thống kê, vì hoàn toàn không có bằng chứng nào khác cho thấy bà giết bệnh nhân. Họ lí luận rằng tất cả 4 trường hợp tử vong de Berk đều có mặt. Họ tính xác suất như thế là 1 trên 342 triệu. Bản thân bà luôn kêu vô tội. Gần đây, một nhóm chuyên gia thống kê Hà Lan chất vấn cách tính và lí luận này. Tòa án sẽ mở phiên xét xử phúc thẩm để xem bà có tôi hay không. Các trường hợp này cho thấy hiểu sai xác suất nhiều khi dẫn đến hệ quả rất nghiêm trọng.