Y HỌC THỰC CHỨNG VÀ ĐẠI DỊCH COVID-19

*Nguyễn Văn Tuấn**

TÓM TẮT:

Đại dịch Covid-19 cùng những vấn đề liên quan đã là đề tài của hàng vạn nghiên cứu trên thế giới. Nhưng bên cạnh một số rất ít nghiên cứu có phẩm chất tốt và tầm ảnh hưởng cao, đa số nghiên cứu không có đóng góp gì quan trọng, thậm chí gây nhiễu y văn. Rất nhiều nghiên cứu có vấn đề về phương pháp và phương pháp luận, kể cả thiết kế, phân tích và diễn giải. Trong bối cảnh như thế, các nguyên lý của Y học Thực chứng (Evidence Based Medicine) có thể giúp cho việc đánh giá nghiên cứu và diễn giải kết quả nghiên cứu đầy đủ hơn. Trong bài này, tôi sẽ bàn qua việc áp dụng các nguyên lý này trong việc thẩm định nghiên cứu Covid-19. Các nghiên cứu Covid-19 sẽ được dùng để minh hoạ cho các khái niệm sai lệch (bias), độ nhạy của nghiên cứu, ý nghĩa trị số P, kiểm định nhiều giả thuyết và phát hiện dương tính giả, và ứng dụng phương pháp Bayes trong diễn giải kết quả nghiên cứu trái ngược nhau. Những tiêu chí đề ra trong bài này sẽ giúp cho việc đánh giá một công trình nghiên cứu có hệ thống hơn, và giúp rút ngắn hơn khoảng cách giữa nghiên cứu khoa học và ứng dụng lâm sàng.

1. ĐẠI DỊCH COVID-19 VÀ NGHIÊN CỨU KHOA HỌC

Một cách chánh thức, dịch Covid-19 bộc phát vào cuối tháng 12 năm 2019. Tuy nhiên, chứng cứ khoa học cho thấy dịch có thể đã xảy ra trước đó 3 tháng. Số liệu từ Vũ Hán cho thấy bệnh nhân đầu tiên được nhập viện là tháng 11/2019. Nguyên nhân của đại dịch là một virus trong chủng *coronavirus* và được định danh là SARS-CoV-2 (*severe acute respiratory syndrome coronavirus* 2). Thoạt đầu, dịch Covid-19 được giới khoa học đánh giá là *epidemic,* nhưng sau một thời gian theo dõi qui mô và tác động của dịch, ngày 12/3/2020 Tổ chức Y tế Thế giới (WHO) chánh thức tuyên bố *pandemic* (đại dịch).

Đại dịch Covid-19 và những vấn đề liên quan trở thành một đề tài cho hàng vạn nghiên cứu khoa học trên thế giới. Như là một qui luật, bất cứ đại dịch mới nào cũng đặt ra nhiều câu hỏi khoa học. Trong đại dịch Covid-19, chúng ta chưa biết rõ cơ chế bệnh sinh, chưa biết đường lan truyền của virus, và ngay cả yếu tố nguy cơ cũng chưa rõ ràng. Bên cạnh đó, y học chưa có thuốc đặc trị và cũng chưa có vaccine phòng ngừa. Đối với nhà chức trách, vấn đề dự báo sự lan toả và ảnh hưởng đến kinh tế của dịch đặt ra nhiều thách thức cho các nghiên cứu dịch tễ học định lượng. Do đó, không ngạc nhiên khi có hàng vạn nhóm nghiên cứu và 'tập đoàn khoa học' trên thế giới tập trung vào nghiên cứu về Covid-19. Ngay cả những labo không có kinh nghiệm về các bệnh truyền nhiễm cũng chuyển sang nghiên cứu Covid-19 với hi vọng đóng góp một phần vào việc chinh phục đại dịch.

Sự hợp tác giữa các nhóm nghiên cứu trên thế giới thật là ngoạn mục và đã cung cấp một lượng thông tin phong phú và nhanh chưa từng có trong lịch sử khoa học. Theo *Science*, nếu chỉ tính từ đầu tháng 1/2020 đến giữa tháng 5/2020, đã có 23,000 bài báo khoa học liên quan đến Covid-19 được công bố, và cứ mỗi 20 ngày, con số này tăng gấp đôi.[1] Vẫn theo *Science*, chỉ trong đầu tuần tháng 5/2020, số bài báo khoa học liên quan đến Covid-19 lên đến 4000[1]!

Đa số (trên 60%) các bài báo công bố trên những thư khố khoa học như *bioRxiv*, *medRxiv*, *ChemRxiv* và *ChinaXiv;*[2] Thư khố bioRxiv cho biết họ nhận bài báo về dịch Covid-19 đầu tiên vào ngày 19/1/2020, nhưng đến tháng 2/2020 thì con số lên đến 281 bài, và đến giữa tháng 4/2020 con số đã hơn 6000! Thư khố là một hình thức công bố do chuyên ngành vật lý khởi xướng hơn 30 năm trước, nhưng chỉ mới phổ biến trong chuyên ngành y sinh học trong thời gian gần đây. Theo mô thức này, tác giả chỉ đơn giản tải bản thảo bài báo (và có khi cả dữ liệu) lên một thư khố. Bản thảo có thể chưa hoàn chỉnh, và cũng chưa qua bình duyệt. Mục tiêu của việc công bố kết quả trên các thư khố là tạo điều kiện để kết quả nghiên cứu đến với cộng đồng khoa học nhanh, và để các nhà khoa học khác có thể bình luận trước khi bài báo được nộp cho một tập san có bình duyệt.

Tuy nhiên, một số khác được công bố trên những tập san y sinh học "chánh thống", hiểu theo nghĩa thuộc các hiệp hội y khoa và xuất bản

* *GS Viện Nghiên cứu Y khoa Garvan; Trường Y St Vincent's, Đại học New South Wales, Australia*

bởi các nhà xuất bản có uy tín cao. Giới y khoa trên thế giới chứng kiến nhiều bài báo liên quan đến Covid-19 được công bố trên những tập san có tầm ảnh hưởng cao (như *New England Journal of Medicine, Lancet, JAMA, BMJ*) có thể nói là từng phút!

Như là một qui luật chung, số lượng nghiên cứu có tương quan nghịch với phẩm chất nghiên cứu. Giới nghiên cứu ước tính rằng khoảng 85% các nghiên cứu y khoa là phung phí, vì tác giả đặt câu hỏi nghiên cứu sai, vì thiết kế nghiên cứu không đúng, vì phân tích sai và diễn giải chủ quan.[3] Tính đến nay đã có hơn 1000 thử nghiệm lâm sàng được đăng ký với Thư viện Quốc gia về Y khoa của Hoa Kỳ. Nhưng theo một số tác giả, các nghiên cứu đó được thiết kế không đạt chuẩn hay cỡ mẫu quá nhỏ để có thể cung cấp thông tin có ích. Một ví dụ tiêu biểu là có đến 145 thử nghiệm về thuốc hydroxychloroquine, nhưng 32 trong số này có cỡ mẫu dưới 100 người, 10 nghiên cứu không có nhóm chứng, và 12 nghiên cứu có mục tiêu so sánh nhưng lại không dùng mô hình ngẫu nhiên hoá.[3]

Theo nhận định của Tập san *BMJ*, những công trình và bài báo được công bố trên những tập san 'chánh thống' hàng đầu trên thế giới, nếu bình thường chỉ có thể công bố trên những tập san "dỏm".[4] Ngay cả thời gian từ lúc nộp bản thảo đến lúc công bố thường chỉ 1 tuần (trong khi đó trong điều kiện bình thường phải mất từ 6 tháng đến 12 tháng), chứng tỏ quá trình bình duyệt các nghiên cứu liên quan đến Covid-19 có vấn đề.

Một ví dụ tiêu biểu về vấn đề trong quá trình bình duyệt là một lá thư trên *New England Journal of Medicine* vào tháng 3/2020.[5] Trong thư, một nhóm nghiên cứu bên Đức cho rằng SARS-CoV-2 có thể lan truyền từ người sang người mà không có triệu chứng. Ngay cả BS Anthony Fauci (cố vấn y tế cho Chánh phủ Trump) cũng cho rằng đây là chứng cớ thuyết phục về sự lây truyền của virus từ người sang người. Nhưng vài ngày sau thì một nhóm nghiên cứu cũng từ Đức công bố một lá thư khác chỉ ra cái sai lầm của lá thư đầu. Trong thực tế, bệnh nhân lây truyền đã có triệu chứng. Tuy lá thư này[5] không rút lại, nhưng kết luận của lá thư thì hoàn toàn sai.

Tình trạng bình duyệt không thấu đáo, vội vã công bố kết quả nghiên cứu dẫn đến hiện tượng bất tái lập (irreproducibility) trong nghiên cứu và một số bài báo khoa học phải bị rút xuống. Chẳng hạn như nghiên cứu về hiệu quả của steroids trong việc điều trị bệnh nhân Covid-19 công bố trên *New England Journal of Medicine* cho thấy dexamethasone giảm nguy cơ tử vong 17% (tỉ số nguy cơ 0.83; khoảng tin cậy 95% dao động từ 0.75 đến 0.93),[6] nhưng chỉ 1 tháng sau, một thử nghiệm lâm sàng khác cho thấy thuốc không có hiệu quả giảm tử vong.[7] Một ví dụ tiêu biểu khác là nghiên cứu quan sát chỉ ra rằng những người có nhóm máu O có nguy cơ nhiễm SARS-Cov-2 thấp hơn 35% so với những người thuộc nhóm máu khác,[8] nhưng một nghiên cứu sau đó cho thấy nhóm máu không có liên quan đến nguy cơ nhiễm SARS-Cov-2.[9] Những kết quả thiếu nhất quán như trên dẫn đến một số bài báo bị thu hồi, và số lượng càng ngày càng nhiều. Tính đến nay (30/8/2020), theo *Retraction Watch*, đã có 32 bài báo khoa học liên quan đến Covid-19 đã bị rút lại (retractionwatch.com/retracted-coronavirus-covid-19-papers). Đó là một con số kỉ lục trong một thời gian rất ngắn, và đặt ra nhiều câu hỏi làm cách nào để nâng cao phẩm chất nghiên cứu y học.

2. Y HỌC THỰC CHỨNG VÀ ĐÁNH GIÁ CHỨNG CỨ KHOA HỌC

Y học thực chứng là một phương pháp thực hành y khoa dựa vào các chứng cứ khoa học một cách sáng suốt và có ý thức, cùng với kĩ năng lâm sàng, nhằm nâng cao chất lượng chăm sóc bệnh nhân.[10] Các chứng cứ là kết quả nghiên cứu y khoa đã được công bố trên các tập san y học chuyên môn. Sử dụng chứng cứ một cách "sáng suốt và có ý thức" có thể hiểu là người thầy thuốc phải cân nhắc, đánh giá, phân loại các dữ liệu nghiên cứu y học, kết hợp cùng kinh nghiệm lâm sàng và thông tin từ bệnh nhân. Tóm lại, 'giáo lý' căn bản của y học thực chứng là bác sĩ cùng làm việc với bệnh nhân, trang bị bằng các dữ liệu khoa học, để đi đến một quyết định, một sự lựa chọn tối ưu cho bệnh nhân.

Do đó, thực hành y học thực chứng bắt đầu bằng việc thẩm định giá trị khoa học của chứng cứ và kết thúc bằng áp dụng chứng cứ trong lâm

Bảng 1: Những tiêu chí chính để đánh giá một nghiên cứu y khoa

Tiêu chí	Chỉ số cụ thể (ví dụ)
Mô hình nghiên cứu	
Mô hình thiết kế là gì	Xem danh sách mô hình nghiên cứu (vd cắt ngang, đoàn hệ, bệnh-chứng, thử nghiệm lâm sàng, v.v.)
Có nhóm chứng	Có / Không
Cỡ mẫu	Xem phần phương pháp và giả định về cách tính
Ngẫu nhiên hoá	Có / Không
Đánh giá kết cục 'blinded'	Có / Không
Phương pháp phân tích đúng	Xem phần phương pháp
Kết quả / chứng cứ	
Mức độ ảnh hưởng	Khoảng tin cậy 95%
Do yếu tố ngẫu nhiên?	Trị số P
Do sai lệch?	Xem xét yếu tố nào có thể ảnh hưởng đến can thiệp và kết cục (outcome)
Do yếu tố nhiễu?	Xem xét yếu tố nào có thể ảnh hưởng đến kết cục, và tác giả có hiệu chỉnh trong phân tích
Do hậu quả đến nguyên nhân	Có phải kết quả báo cáo là do mối liên quan nghịch đảo, tức hậu quả dẫn đến nguyên nhân
Do nguyên nhân đến hậu quả	Nếu các yếu tố ngẫu nhiên, sai lệch, nhiễu bị loại bỏ, kết quả có thể là do mối liên hệ nhân quả

sàng. Có thể đề ra 2 tiêu chí để thẩm định một nghiên cứu: đánh giá mô hình nghiên cứu và đánh giá chứng cứ thực tế (Bảng 1).

Dĩ nhiên, nguyên lý học thực chứng còn tiêu chí thứ ba về áp dụng chứng cứ. Sau khi thậm định giá trị của kết quả một nghiên cứu, câu hỏi kế tiếp là kết quả này có thể áp dụng trong thực tế. Để trả lời câu hỏi này, người thầy thuốc phải so sánh đặc điểm của đối tượng nghiên cứu (Giới tính, tuổi, tiền sử lâm sàng, v.v.) có giống như bệnh nhân đang được điều trị; bệnh viện có phương tiện xét nghiệm hay đo lường như mô tả trong công trình nghiên cứu; phương pháp can thiệp có sẵn, chuyên gia có kinh nghiệm có sẵn, chi phí can thiệp có thể chấp nhận được hay không; và Xem xét và cân bằng giữa lợi ích và tác hại (có thể).

2.1 Đánh giá mô hình nghiên cứu

Chứng cứ khoa học được đúc kết từ những công trình nghiên cứu. Các công trình nghiên cứu thường được thiết kế theo nhiều mô hình khác nhau, và giá trị khoa học của các mô hình nghiên cứu cũng khác nhau. Theo hệ thống phân loại của y học thực chứng, giá trị của chứng cứ khoa học được đánh giá theo thứ tự, cao nhất đến thấp nhất như sau (xem Hình 1):

Các phân tích tổng hợp (meta-analysis) và tổng quan có hệ thống các công trình thử nghiệm lâm sàng có đối chứng ngẫu nhiên (RCT hay randomized controlled trials).

Các công trình thử nghiệm RCT riêng lẻ. Đây là các công trình nghiên cứu có số đối tượng lên đến hàng ngàn, thậm chí hàng vạn, để làm tiêu chuẩn vàng cho việc đánh giá mức độ an toàn và hiệu quả của một can thiệp.

Các nghiên cứu đoàn hệ, theo dõi đối tượng nghiên cứu theo thời gian (cohort studies). Đây là các nghiên cứu quan sát (không can thiệp), có theo dõi đối tượng theo thời gian, và thường có nhóm chứng để so sánh. Nhưng vì các nhóm đối tượng hình thành 'tự nhiên' (không có chia nhóm ngẫu nhiên), nên kết quả thường chịu ảnh hưởng của thiên lệch và yếu tố nhiễu.

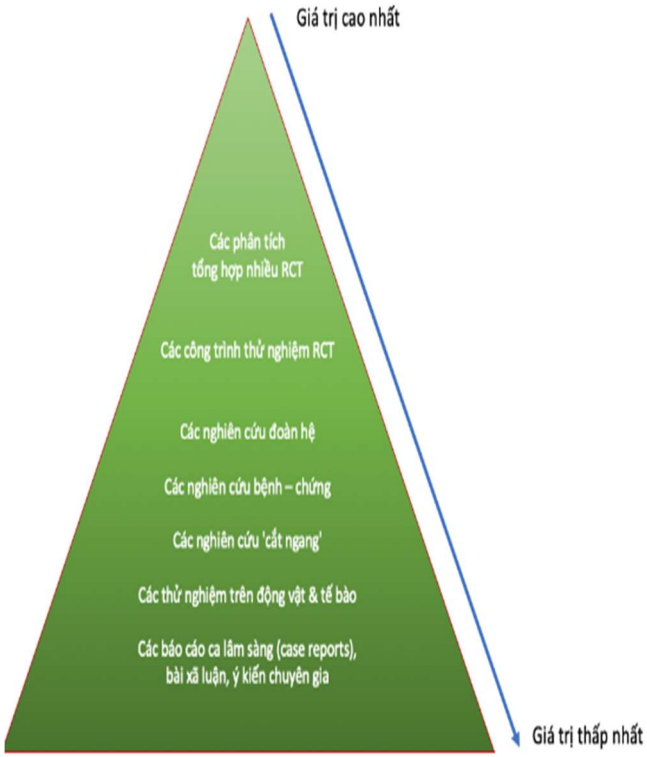
Các nghiên cứu bệnh – chứng (case-control studies). Các nghiên cứu theo mô hình này cũng không có can thiệp theo thời gian, nhưng có nhóm chứng. Mục tiêu chủ yếu của nghiên cứu bệnh chứng là nhằm xác định và đánh giá yếu tố nguy cơ của một bệnh lý. Vì nghiên cứu theo mô hình bệnh chứng phải truy cập dữ liệu ngược thời gian (retrospective) nên kết quả thường bị sai lệch nếu nhóm chứng không được chọn cẩn thận.

Các nghiên cứu 'cắt ngang' (cross-sectional studies). Các nghiên cứu theo mô hình này chỉ đo lường đối tượng nghiên cứu tại một thời điểm, thường có mục tiêu ước tính qui mô bệnh qua tỉ lệ hiện hành (prevalence).

Các thử nghiệm trên động vật và dòng tế bào. Đây là các nghiên cứu sơ khởi để kiểm định một giả thuyết khoa học, thường được thiết kế chặt chẽ, có nhóm chứng, và có theo dõi các đối tượng theo thời gian. Nhưng vì là nghiên cứu cơ bản, nên khi áp dụng trên người thì đa số kết quả không mang tính tái lập hoặc thất bại.

Các báo cáo ca lâm sàng (case reports), bài xã luận, ý kiến chuyên gia. Các ca lâm sàng là những báo cáo rất có ích cho kinh nghiệm lâm sàng, nhưng không giúp tìm ra một qui luật hay

một mối liên quan. Ý kiến chuyên gia tuy quan trọng, nhưng thường hay chịu sự chi phối của yếu tố cá nhân và mâu thuẫn lợi ích.



Hình 1: Phân biệt giá trị khoa học của các mô hình nghiên cứu y học

Những gì xảy ra cho thấy một lần nữa rằng các nghiên cứu quan sát, dù là có can thiệp, nhưng nếu không có nhóm chứng thì rất khó diễn giải. Đó cũng chính là lý do tại sao các nghiên cứu quan sát có can thiệp hay không can thiệp không thể có giá trị khoa học cao như nghiên cứu RCT.

Ngoài việc đánh giá qua mô hình nghiên cứu theo hình tháp trên, giá trị khoa học của một nghiên cứu còn được xem xét qua chi tiết về phương pháp nghiên cứu. Những chi tiết này bao gồm những vấn đề liên quan đến thiết kế như ước tính cỡ mẫu, các yếu tố thiên lệch và yếu tố nhiễu, và phương pháp phân tích.

2.2 Cỡ mẫu

Giáo sư Karl Pearson, triết gia và người sáng lập ra Khoa học Thống kê hiện đại, từng nói rằng "Sự hữu dụng của khoa học chỉ ở phương pháp, chứ không phải chất liệu".[13] Mặc dù một phương pháp có thể áp dụng cho nhiều nghiên cứu khác nhau, nhưng chi tiết về phương pháp luận và phương pháp đo lường và phân tích mới chính là những yếu tố quyết định phẩm chất khoa học của một công trình nghiên cứu. Các

Trường hợp 1: Hydroxychloroquine (HCQ) and azithromycin.

Gautret và đồng nghiệp[11] công bố kết quả nghiên cứu về hiệu quả của HCQ và azithromycin trong điều trị các bệnh nhân nhiễm SARS-cov-2. Đây là một nghiên cứu theo dạng ca lâm sàng. Nhóm nghiên cứu tuyển chọn 80 bệnh nhân được xác định là nhiễm SARS-cov-2 vào công trình nghiên cứu. Trong số 80 người, gần phân nửa (42) là nam giới. Tuổi trung bình là 52 (min - max: 18 đến 88). Thời gian từ lúc có triệu chứng đến nhập viện là ~5 ngày (min - max: 1 - 17). Tất cả điều được điều trị bằng HCQ và azithromycin ít nhứt là 3 ngày. Bệnh nhân được theo dõi 14 ngày. Tiêu chí lâm sàng là: (a) số bệnh nhân cần oxygen trị liệu hoặc chuyển sang ICU sau 3 ngày điều trị; (b) mức độ truyền nhiễm qua xét nghiệm PCR; và (c) thời gian nằm viện. Kết quả có thể tóm lược như sau: dựa trên tiêu chí lâm sàng đề ra, tác giả báo cáo rằng có 65 bệnh nhân (81%) 'đạt'. Chỉ có 15% cần oxygen trị liệu; 3 bệnh nhân chuyển sang ICU (và trong số này 2 người về lại khoa truyền nhiễm). Có 1 bệnh nhân 86 tuổi tử vong. Đánh giá bằng qPCR cho thấy số ca âm tính giảm hẳn: ngày 7 có 83%; ngày 8 là 93%.

Tuy nhiên, nghiên cứu này thu hút rất nhiều phê bình gay gắt từ các chuyên gia bệnh truyền nhiễm trên thế giới. Ngoài vấn đề tác giả sử dụng sai phương pháp trong phân tích dữ liệu, các nhà nghiên cứu chỉ ra rằng đây là nghiên cứu không có nhóm chứng. Vì thiếu nhóm chứng để so sánh, nên những kết quả tác giả trình bày rất ư khó diễn giải.

Khoảng 4 tháng sau, một công trình thử nghiệm RCT được công bố trên Tập san *New England Journal of Medicine,*[12] nhưng kết quả thì hoàn toàn khác với kết luận của Gautret và đồng nghiệp.[11] Nghiên cứu RCT trên 667 bệnh nhân, được chia ngẫu nhiên thành 3 nhóm: nhóm được điều trị theo phác đồ hiện hành (nhóm chứng), nhóm được cho dùng HCQ 400 mg 2 lần / ngày, và nhóm dùng HQC 400 mg 2 lần/ngày cộng với azithromycin 500 mg / ngày. Thời gian điều trị là 7 ngày. Sau 15 ngày theo dõi, các nhà nghiên cứu không phát hiện khác biệt về các chỉ số lâm sàng giữa 3 nhóm. Các tác giả kết luận rằng ở những bệnh nhân nhập viện vì Covid-19 trong tình trạng nhẹ đến trung, HCQ hoặc HCQ + azithromycin không tốt hơn so với phác đồ điều trị hiện hành.[12]

chi tiết về thiết kế và phân tích có thể tìm trong các bản hướng dẫn như CONSORT[14] cho các nghiên cứu RCT, STROBE[15] cho các nghiên cứu quan sát, và ARRIVE[16] cho các nghiên cứu liên quan đến động vật và tế bào.

Một trong những chi tiết quan trọng của các bản hướng dẫn này là ước tính cỡ mẫu. Như là một qui luật, các nghiên cứu với cỡ mẫu không đủ hay quá ít sẽ có độ nhạy (power) thấp để phát hiện một mối liên quan. Nhiều nhà khoa học không nhận ra rằng các nghiên cứu với độ nhạy thấp (power thấp hơn 80% chẳng hạn) thường cho ra những kết quả rất lạc quan và có ý nghĩa thống kê, nhưng đó có thể là những kết quả dương tính giả và có khả năng tái lập (reproducibility) rất thấp.[17]

> ***Trường hợp 2:*** Nghiên cứu hiệu quả của remdesivir trong điều trị bệnh nhân nhiễm SARS-Cov-2
>
> Wang và đồng nghiệp thực hiện một thử nghiệm RCT[19] trên 237 bệnh nhân nhiễm SARS-Cov-2, trong đó 158 người được điều trị bằng remdesivir và 79 người trong nhóm chứng (placebo). Tỉ lệ bệnh nhân bình phục (sau 14 ngày) là 27% ở nhóm can thiệp và 23% ở nhóm chứng. Mức độ khác biệt không có ý nghĩa thống kê (khoảng tin cậy 95% dao động từ -8.1 đến 15.1 ngày).
>
> Trong phần phương pháp, tác giả cho biết nghiên cứu cần quan sát 325 biến cố của 2 nhóm để có độ nhạy (power) 80% và sai số loại I là 2.5%. Thế nhưng trong thực tế, nghiên cứu quan sát được 148 ca bình phục sau 28 ngày. Có thể nói đây là một nghiên cứu có kết quả 'âm tính', nhưng có thể lý do là số cỡ mẫu (trong trường hợp này là số biến cố) không đủ.

Phương pháp ước tính cỡ phụ thuộc vào mô hình nghiên cứu, đặc điểm của biến phụ thuộc, giả định về mức độ ảnh hưởng. Không có một công thức ước tính cỡ mẫu cố định. Do đó, mô tả đầy đủ về phương pháp ước tính cỡ mẫu, như đặc điểm của biến phụ thuộc, mức độ ảnh hưởng, độ dao động, sai sót loại I và sai sót loại II rất cần thiết để độc giả có thể đánh giá độ tin cậy của nghiên cứu.[18] Tuy nhiên, rất tiếc là tuyệt đại đa số các nghiên cứu liên quan đến Covid-19 không báo cáo các chi tiết đó. Do đó, độc giả không biết tại sao tác giả đi đến cỡ mẫu mà họ báo cáo. Ngoài ra, đa số các nghiên cứu trên chuột thường có số cỡ mẫu rất thấp, và tác giả không giải thích về số cỡ mẫu đó. Do đó, người đọc rất khó đánh giá kết quả trong bối cảnh cỡ mẫu và những giả định đằng sau của công trình nghiên cứu.

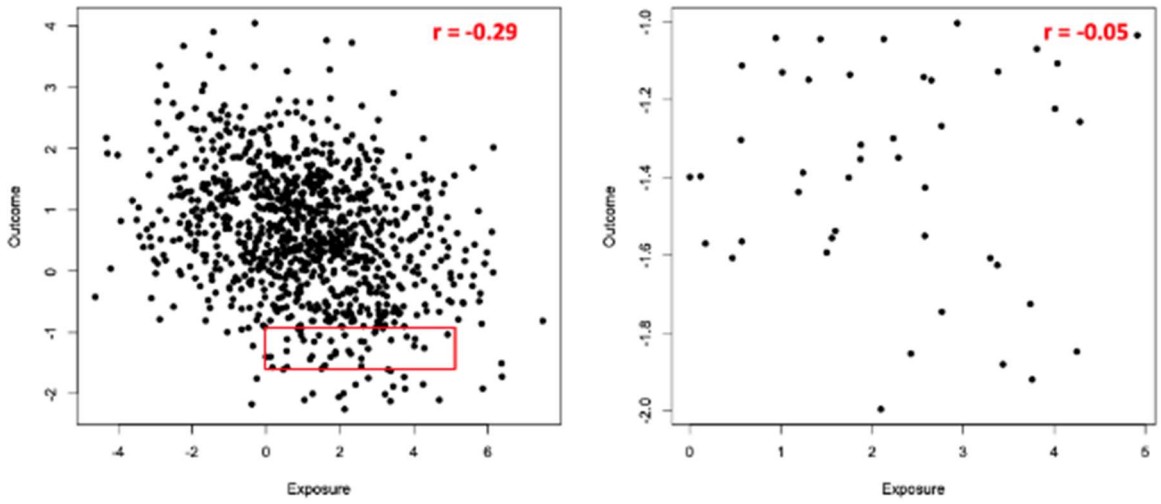
2.3 'Biases' và yếu tố nhiễu

Kết quả của các nghiên cứu can thiệp không có nhóm chứng và nghiên cứu quan sát thường chịu sự chi phối của các yếu tố *bias* (tôi tạm dịch là *thiên lệch*) và *confounders* (*yếu tố nhiễu*). Thiên lệch xảy ra khi kết quả quan sát được khác hay lệch so với giá trị thật. Nếu giá trị thật của tỉ số odds (OR) là 1.0, nhưng kết quả nghiên cứu cho ra OR 1.2, thì đó là *bias*. Yếu tố nhiễu được định nghĩa là có ảnh hưởng đến cả yếu tố nguy cơ (risk factor) và biến phụ thuộc (outcome). Một dạng 'nhiễu' khác có tên là *collider* và tôi tạm dịch là *sai lệch đồng căn.* Khi cả yếu tố nguy cơ và biến phụ thuộc có ảnh hưởng đến biến *Z*, thì Z được gọi là *collider* hay *sai lệch đồng căn.*

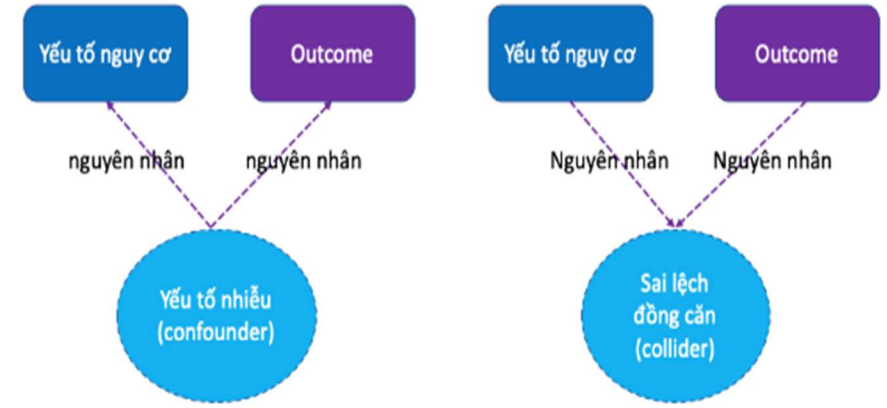
Danh sách về thiên lệch rất dài,[20] và những thiên lệch này gần như chỉ xảy ra ở các nghiên cứu quan sát và thiếu nhóm chứng. Trong danh sách đó, *thiên lệch lựa chọn* (*selection bias*) là một mối đe doạ thường xuyên cho nghiên cứu khoa học. Thiên lệch lựa chọn xảy ra trong các nghiên cứu khi đối tượng được chọn từ một mẫu khác biệt với quần thể (population). Vì sự khác biệt, mối liên quan quan sát trong mẫu nghiên cứu có thể rất lệch so với mối liên quan trong quần thể.

Hình 2 minh hoạ một trường hợp giả tưởng về mối liên quan giữa yếu tố phơi nhiễm và biến phụ thuộc trong một quần thể, với hệ số tương quan r = -0.29 (P < 0.0001; biểu đồ bên trái). Tuy nhiên, nếu một mẫu nhỏ được tuyển chọn từ quần thể đó được chọn (biểu dồ bên phải) thì mối liên quan không còn có ý nghĩa thống kê (r = -0.05; P = 0.72). Do đó, những nghiên cứu trên một nhóm nhỏ, được chọn một cách thiên lệch từ quần thể, có nguy cơ cao cho kết luận sai.

Sai lệch đồng căn[21] là một đe doạ đối với các nghiên cứu quan sát. Cần phải nhấn mạnh rằng sai lệch đồng căn khác với yếu tố nhiễu (Hình 3). Giả dụ như trong thực tế, không có mối liên quan giữa hút thuốc lá và nhiễm Covid-19. Để



Hình 2: Mối liên quan giữa exposure và outcome hiện hữu trong quần thể (hình trái), nhưng nếu một nhóm nhỏ được chọn không đại diện từ quần thể có thể không phát hiện được mối liên quan (hình phải).



→ Hình 3: Yếu tố nhiễu là yếu tố có ảnh hưởng [nhân quả] đến yếu tố nguy cơ và outcome (hình trái), nhưng yếu tố sai lệch đồng căn thì chịu sự ảnh hưởng của cả yếu tố nguy cơ và outcome. Cả hai yếu tố nhiễu và sai lệch đồng căn đều có thể làm sai lệch mối liên quan nhân quả giữa yếu tố nguy cơ và outcome, nếu không được hiệu chỉnh đúng.

tìm hiểu xem có thể quan sát một mối liên quan giữa hút thuốc lá và nhiễm virus do vấn đề chọn mẫu, các nhà nghiên cứu dùng mô phỏng. Kết quả mô phỏng cho thấy nhà nghiên cứu có thể quan sát hút thuốc lá tăng nguy cơ nhiễm gấp 2 lần, hoặc giảm nguy cơ lây nhiễm gần 50%[22]! Có thể giải thích hiện tượng này như sau: thoạt đầu, đa số người được xét nghiệm là nhân viên y tế, và đa số nhân viên y tế không hút thuốc lá. Nói cách khác, những người hút thuốc là bị 'loại bỏ' khỏi chương trình xét nghiệm. Hệ quả là trong số những người được xét nghiệm có nhiều người không hút thuốc lá, làm cho phân tích thống kê phát hiện không hút thuốc lá tăng nguy cơ nhiễm virus! Tóm lại, khi đọc một nghiên cứu quan sát (không can thiệp), chúng ta cần phải cảnh giác với các loại thiên lệch, đặc biệt là sai lệch đồng căn. Những mối liên quan giữa yếu tố phơi nhiễm và biến phụ thuộc có thể không phải là thật mà do ảnh hưởng của yếu

Trường hợp 3: Sai lệch đồng căn và đánh giá yếu tố nguy cơ COVID-19

Để đánh giá ảnh hưởng của cách chọn mẫu có ảnh hưởng đến khả năng xét nghiệm Covi-19 (không cần biết dương tính hay âm tính), các nhà nghiên cứu dùng dữ liệu của UKBiobank (có khoảng 500,000 đối tượng nghiên cứu). Họ phân tích 2556 yếu tố và phát hiện có 32% trong số các yếu tố đó có tham gia xét nghiệm. Một trong những mối liên quan đó là giữa ACE2 và tử vong ở bệnh nhân Covid-19, rất có thể cũng chỉ do sai lệch đồng căn, chứ không phải mối liên hệ nhân quả. Nguồn: Griffith G, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. MedRxiv 8/5/2020. https://doi.org/10.1101/2020.05.04.20090506

tố sai lệch đồng căn. Nếu không là yếu tố sai lệch đồng căn, thì phải nghĩ ngay đến yếu tố nhiễu. Nếu cả hai yếu tố đều loại trừ, thì lúc đó chúng ta mới đặt niềm tin vào kết quả nghiên cứu.

Các mô hình hồi qui là một phương pháp tốt để hiệu chỉnh cho các yếu tố thiên lệch, và giúp suy luận về mối liên hệ nhân quả chính xác hơn. Tuy nhiên, một số nhà khoa học có xu hướng hiệu chỉnh cho tất cả các biến số, với ý định sẽ giảm thiểu tối đa sự ảnh hưởng của yếu tố thiên lệch. Chẳng hạn như một số tác giả có xu hướng hiệu chỉnh cho các yếu tố như cân nặng, chiều cao, tỉ trọng cơ thể (body mass index) Và độ tuổi trong một mô hình hồi qui logistic. Nhưng cách phân tích đó phản tác dụng, bởi vì mô hình hiệu chỉnh sẽ dẫn đến tình trạng hiệu chỉnh thái quá (*over-adjustment*) và *over-fitting*. Đó là chưa đề cập đến tình trạng đa cộng tuyến (các yếu tố nguy cơ có liên quan với nhau).

Do đó, *hiệu chỉnh cho yếu tố nhiễu đòi hỏi một sự cân nhắc và xem xét cẩn thận dựa trên kiến thức chuyên ngành, chứ không phải thống kê.* Cũng cần phải nhắc lại rằng không phải mối liên quan nào cũng cần phải hiệu chỉnh. Chẳng hạn như nếu một yếu tố không ảnh hưởng đến mối liên quan giữa yếu tố nguy cơ và biến phụ thuộc thì không cần phải hiệu chỉnh.[23]

2.4 Phương pháp phân tích

Phương pháp phân tích dữ liệu tuỳ thuộc vào mô hình nghiên cứu và bản chất của dữ liệu. Chẳng hạn như các nghiên cứu theo thời gian, đối tượng được theo dõi và đo lường nhiều lần, rất có ích để nhà nghiên cứu có thể đánh giá sự biến chuyển của các yếu tố bệnh sinh theo thời gian. Nhưng phân tích dữ liệu từ các nghiên cứu thời gian đặt ra nhiều thách thức mang tính phương pháp luận. Lý do là vì: (i) các đo lường trong mỗi cá nhân có liên quan với nhau; (ii) khoảng cách giữa 2 đo lường khác nhau giữa các bệnh nhân; và (iii) thường có giá trị trống (missing data).

Một số tác giả dùng phương pháp phân tích phương sai (ANOVA) để phân tích dữ liệu từ nghiên cứu theo thời gian. Tuy nhiên, phương pháp ANOVA đơn giản không thích hợp cho nghiên cứu theo thời gian và xử lý dữ liệu trống. Nếu sự tương quan trong mỗi đối tượng không được hiệu chỉnh, thì kết quả phân tích có thể sai lệch, thậm chí cho ra kết quả có ý nghĩa thống kê nhưng là dương tính giả.[24]

Phân tích dữ liệu từ nghiên cứu theo thời gian cần đến những phương pháp 'hiện đại' như GEE (generalized estimating equations)[25] mô hình ảnh hưởng hỗn hợp (linear mixed effects model).[26] Một lợi điểm của các phương pháp này là có thể xử lý các giá trị trống mà vẫn có thể hiệu chỉnh cho những dao động và tương quan trong mỗi đối tượng nghiên cứu.

> ***Trường hợp 4:*** Nghiên cứu về hiệu quả của HCQ và azithromycin trong điều trị các bệnh nhân nhiễm SARS-cov-2 do Gautret và đồng nghiệp.[11] thực hiện theo dõi kết quả xét nghiệm RT-PCR ngày 2, 4 và 6. Sau đó, họ phân tích sự khác biệt về tỉ lệ âm tính RT-PCR giữa 2 bệnh nhân từng ngày. Đây là một sai sót tiêu biểu trong phân tích dữ liệu theo thời gian, bởi vì phương pháp phân tích từng ngày không phản ảnh được sự dao động trong *mỗi bệnh nhân,* và dễ dẫn đến kết quả sai lệch. Khi chúng tôi phân tích lại dữ liệu bằng mô hình ảnh hưởng hỗn hợp thì kết quả cho thấy thuốc HCQ không có hiệu quả giảm tỉ lệ RT-PCR (P = 0.355) như tác giả kết luận.[27]

Một vấn đề phổ biến liên quan đến phân tích dữ liệu theo thời gian (như thí nghiệm trước – sau) là cách tính phần trăm biến đổi. Chẳng hạn như một nghiên cứu đo lường chỉ số nhiễm trước (ký hiệu là x_0) và sau (x_1) khi can thiệp, đa số các nhà khoa học tính phần trăm biến đổi bằng cách lấy hiệu số chia cho giá trị trước khi can thiệp theo công thức

$$(x_1 - x_0)/x_0 \times 100$$

và dùng số phần trăm đó cho các phân tích tiếp theo. Mặc dù đây là cách tính rất phổ biến, nhưng cách tính đó sai vì sự mất cân đối[28] và có thể dẫn đến kết quả sai lệch.[29] Nếu một cá nhân có giá trị đo lường ban đầu là 1.0 và đo lường lần 2 là 0.80, phương pháp trên sẽ cho ra kết quả giảm 20%; nhưng nếu lấy giá trị ban đầu chia cho giá trị lần 2 thì tăng 25%! Một cách tính cân đối nên dùng số trung bình ở mẫu số:

$$(x_1 - x_0)/trung\ bình\ (x_0, x_1) \times 100$$

Với mẫu số là số trung bình sẽ tránh tình trạng mất cân đối. Trong ví dụ trên, giá trị lần 2 giảm 22%, hay giá trị lần đầu tăng 22% so với giá trị trung bình. Tuy nhiên, một phương pháp tốt hơn nữa là dùng mô hình phân tích hiệp phương sai hay còn gọi là *analysis of covariance.*[30]

Phân nhóm trên biến liên tục

Một trong những 'thói quen' trong phân tích dữ liệu là chia biến liên tục thành nhiều nhóm nhỏ. Chẳng hạn như chia độ tuổi thành từng nhóm tuổi 10-năm hay 5-năm, chia mật độ xương thành 3 nhóm 'bình thường', 'thiếu xương', và 'loãng xương', hay chia tỉ trọng cơ thể thành 4 nhóm 'thiếu cân', 'bình thường', 'quá cân', và 'béo phì'.[31] Nói cách khác, cách làm này biến một biến số liên tục thành một biến phân loại, và dùng biến phân loại cho các phân tích khác. Cách làm này tuy hợp lý trên phương diện *lâm sàng* (chẩn đoán), nhưng hoàn toàn không thích hợp cho mục tiêu *khoa học*.

Trong nghiên cứu khoa học, chia biến liên tục thành 2 hay nhiều nhóm là một cách làm không cần thiết và có thể dẫn đến kết quả sai lệch.[32] Nếu lấy ngưỡng 1.0 để chia nhóm, 2 bệnh nhân với giá trị 0.50 và 1.50 cũng thuộc hai nhóm khác nhau, nhưng 2 bệnh nhân với giá trị 1.05 và 0.98 thuộc về 2 nhóm khác nhau, dù mức độ khác biệt chỉ có 0.07! Ví dụ đơn giản này cho thấy chia biến liên tục dẫn đến hiện tượng *mất thông tin* (loss of information). Ngoài ra, một biến liên tục thường được đo lường với sai số đo lường (measurement error), và việc chia nhóm sẽ dẫn đến sai nhóm. Một cá nhân có thể có giá trị 1.05 hôm nay, nhưng nếu đo một lần nữa có thể là 0.98, và do đó nếu dựa vào đo lường tại một thời điểm thì rất dễ dẫn đến sai lầm trong chia nhóm.

Về phương diện phân tích, tất cả các cá nhân trên hay dưới ngưỡng phân nhóm được xem là tương đương nhau, nhưng trong thực tế giá trị đo lường rất khác nhau, và kết quả cũng khác nhau. Do đó, chia nhóm càng đơn giản (như 2 nhóm) sẽ dẫn đến sự mất thông tin càng nhiều so với chia nhiều nhóm (như 10). Quan trọng hơn nữa, chia nhóm tuỳ tiện làm cho việc hiệu chỉnh bằng mô hình hồi qui trở nên kém hiệu quả. Trong mô hình hồi qui tuyến tính, một biến phân nhóm chỉ giảm mức độ thiên lệch 67% so với một biến liên tục.[33]

Do đó, nếu mục tiêu là nghiên cứu khoa học (không phải lâm sàng), việc chia nhóm từ biến liên tục là không khuyến khích. Trong thực tế, có một số biến liên tục không tuân theo luật phân bố chuẩn (và nhiều người lấy lý do đó biện minh cho việc chia nhóm), nhưng trong tình huống này, các mô hình hồi qui dạng splin hay 'non-parametric smoother' có thể ứng dụng rất hữu hiệu.

Phương pháp chọn yếu tố liên quan

Nhiều nghiên cứu đo lường nhiều hay rất nhiều yếu tố (như $x_1, x_2, x_3, \ldots, x_p$), và câu hỏi nghiên cứu thường là trong số đó, biến số nào có liên quan đến biến phụ thuộc (y). Biến phụ thuộc có thể là biến liên tục hay biến nhị phân. Trong điều kiện có quá nhiều biến tiên lượng (tức p có thể hàng trăm, thậm chí hàng triệu), câu hỏi đó rất quan trọng, và là chủ đề của rất nhiều người cứu trong quá khứ.

Nếu nghiên cứu chỉ có 2 biến tiên lượng hay 2 yếu tố nguy cơ ($p = 2$), thì số mô hình khả dĩ là $2^p = 4$. Nếu nghiên cứu có 30 biến tiên lượng, thì số mô hình khả dĩ lên đến 1,073,741,824. Và, xác định mô hình nào hay yếu tố nào có liên quan đến y là một thách thức lớn và đòi hỏi đầu tư về thời gian và suy nghĩ. Vấn đề càng trở nên phức tạp khi các biến x có liên quan với nhau (hiện tượng đa cộng tuyến). Nhà nghiên cứu đối phó với thách thức này bằng nhiều cách, nhưng 2 cách phổ biến nhất có lẽ là:

- Cách thứ nhất là kiểm tra mối liên quan từng biến một với y, và nhận ra biến nào có ý nghĩa thống kê. Sau đó, những biến có ý nghĩa thống kê sẽ đưa vào mô hình đa biến.
- Cách thứ hai là dùng thuật toán 'stepwise' có sẵn trong các chương trình máy tính (như SPSS, SAS, Stata) để chọn các biến liên quan.

Cả hai cách đều có nhiều vấn đề. Cách thứ nhất về căn bản là phân tích đơn biến, tức đánh giá từng biến một. Do đó, vấn đề có thể xảy ra khi biến x_1 có thể có ý nghĩa thống kê trong mô hình đơn biến, nhưng trở nên không có ý nghĩa thống kê khi có sự hiện diện của x_2. Ngoài ra, sai sót loại I (alpha) trong điều kiện kiểm định nhiều mối liên quan cũng sẽ cao bất thường, dẫn đến tỉ lệ dương tính giả tăng. Do đó, cách làm thứ nhất là rất phi khoa học và dễ dẫn đến nhiều kết quả sai.

Phương pháp stepwise[34] cũng dẫn đến nhiều sai sót. Nhiều nhà nghiên cứu không hay chưa nhận thức rằng phương pháp stepwise ít khi nào cho ra một mô hình tốt nhất, nếu có những biến thừa (redundant predictors). Do đó, những biến số thật sự có liên quan với *ý* nhưng phương pháp

stepwise không phát hiện, vì không đạt ngưỡng trị số P để có ý nghĩa thống kê.[35] Hệ quả là mô hình mà phương pháp stepwise nhận ra thường không có độ tái lập cao.

Phương pháp thích hợp nhất, nếu không muốn nói là 'chuẩn', trong việc nhận dạng các yếu tố liên quan trong mô hình đa biến là Bayesian Model Averaging (BMA)[36,37] hoặc LASSO.[38] Cả hai phương pháp BMA và LASSO đã được chứng minh tốt hơn phương pháp stepwise trong việc tìm mô hình tối ưu, và rất có ích cho dữ liệu lớn.

Tuy nhiên, điều quan trọng cần nhấn mạnh là các phương pháp thống kê là công cụ. Tìm mô hình tối ưu hay yếu tố liên quan đòi hỏi kiến thức chuyên ngành. Kiến thức chuyên ngành bao gồm lâm sàng, sinh học, dịch tễ học, và thực tiễn. Phương pháp thống kê không thể thay thế kiến thức chuyên môn về vấn đề nghiên cứu.

Over-fitting

Over-fitting (có thể hiểu là '*mô hình quá cỡ*') đề cập đến tình trạng dùng mô hình phức tạp để giải thích một liên quan. Chẳng hạn như mối liên quan trong thực tế chỉ cần 2 biến tiên lượng, nhưng nhà nghiên cứu dùng mô hình với hơn 2 biến tiên lượng, và trường hợp đó chính là over-fitting. Có thể ví von hiện tượng over-fitting như người mặc trang phục bó sát người.

Các mô hình đa biến (multivariable model) lúc nào cũng có nguy cơ over-fitting. Đối với các nghiên cứu có nhiều biến số, nhà nghiên cứu thường có xu hướng đưa vào mô hình tất cả các biến số đo lường được. Mô hình như thế rất có thể cho ra giá trị tiên lượng gần với giá trị quan sát, và nhà nghiên cứu có thể hài lòng vì thấy mô hình tốt. Tuy nhiên, khi mô hình đó áp dụng cho một quần thể khác thì hoàn toàn thất bại. Lý do là vì mô hình quá cỡ cố gắng giải thích phần ngẫu nhiên hơn là mối liên quan trong dữ liệu; do đó, mô hình quá cỡ có thể rất tốt cho một nghiên cứu này nhưng hoàn toàn thất bại cho một nghiên cứu độc lập.

Mô hình quá cỡ còn xảy ra khi số biến cố (events) thấp hơn số biến số trong mô hình. Chẳng hạn như nghiên cứu quan sát được 10 ca tử vong, nhưng nhà nghiên cứu phân tích bằng một mô hình có 15 biến tiên lượng, mô hình trong trường hợp này cũng là *quá cỡ*. Trong quá khứ, nhiều nhà khoa học giả định rằng mỗi biến tiên lượng trong mô hình cần phải có ít nhất 10 biến cố;[39] và theo đó, nếu mô hình có 5 biến tiên lượng thì nghiên cứu cần quan sát 50 biến cố. Tuy nhiên, trong vài năm gần đây, giả định này đã được chỉ ra là quá đơn giản và có thể sai. Nghiên cứu lý thuyết cho thấy số biến cố cần thiết cho một mô hình đa biến tuỳ thuộc vào: (i) tỉ lệ phát sinh hay incidence của bệnh, (ii) số yếu tố nguy cơ hay số biến tiên lượng, (iii) phần trăm phương sai mà mô hình giải thích được, và (iv) yếu tố *shrinkage*.[40]

Trường hợp 5: Mô hình tiên lượng dịch COVID-19

Dịch COVOD-19 và những vấn đề liên quan dẫn đến rất nhiều nhóm nghiên cứu trên thế giới xây dựng các mô hình tiên lượng. Các mô hình này tập trung vào việc dự báo diễn biến của dịch, mô hình phát hiện cá nhân có nguy cơ nhiễm cao, tiên lượng nguy cơ tử vong, tiên lượng 'outcome' trong điều trị, v.v. Wyants và đồng nghiệp thực hiện một phân tích tổng quan trên 145 mô hình tiên lượng đã được công bố trên các tập san y khoa,[42] với chỉ số c khá cao (từ 0.73 đến 0.99). Tuy nhiên, tác giả nhận xét rằng tất cả các mô hình tiên lượng đều có yếu tố thiên lệch, bởi vì đối tượng nghiên cứu không mang tính đại diện cho nhóm chứng, hoặc loại bỏ những biến cố lâm sàng quan trọng. Nhiều mô hình có số biến cố thấp hay số biến tiên lượng quá nhiều dẫn đến tình trạng over-fitting hay nguy cơ over-fitting cao. Ngoài ra, đa số mô hình tiên lượng không được kiểm định độ chính xác (calibration). Nhóm tác giả kết luận rằng không một mô hình nào có thể sử dụng trong thực tế.[42]

Các phương pháp hiện đại như LASSO[38] hay ridge regression[41] có thể giúp giảm tình trạng mô hình quá cỡ. Đặc biệt là với phương pháp LASSO, các tham số trong mô hình được 'co' (shrinkage) về 0 bằng cách áp đặt một yếu tố hạn chế tổng số giá trị của ước số. Sự áp đặt này giúp loại bỏ những biến không quan trọng trong mô hình và tránh tình trạng *over-fitting*.

2.5 Diễn giải kết quả nghiên cứu: Yếu tố ngẫu nhiên và trị số P

Có thể nói rằng trong đa số trường hợp, việc diễn giải và kết luận của một công trình nghiên cứu đều qui về trị số P. Kể từ lúc phát kiến vào

thập niên 1920, trị số P đã trở nên vô cùng phổ biến trong khoa học, đến nỗi giới khoa học xem trị số P như là một "*giấy thông hành cho công bố khoa học*."

Rất nhiều nhà khoa học có xu hướng diễn giải trị số P theo mô thức nhị phân, với ngưỡng 0.05 là chuẩn để tuyên bố khám phá. Một kết quả với P < 0.05 được xem là 'có ý nghĩa' (significant), và một kết quả với P > 0.05 'là không có ý nghĩa'. Trong nhiều trường hợp cực đoan, một kết quả với P = 0.04 được cho là có ý nghĩa, còn một kết quả với P = 0.055 thì được diễn giải như là không có liên quan, không có ảnh hưởng. Tuy nhiên, ít người trong giới khoa học nhận ra rằng trị số P dao động rất lớn giữa các mẫu nghiên cứu.[43] Chỉ cần loại bỏ 1 giá trị hay thêm một đối tượng nghiên cứu thì 'ý nghĩa thống kê' có thể thay đổi theo, dù bản chất của mối liên quan không hề thay đổi. Điều này nói lên rằng việc phân định giữa "có ý nghĩa" và "không có ý nghĩa" dựa vào một ngưỡng như 0.05 không nên phải là cách diễn giải tốt. Kết luận về một mối liên quan hay ảnh hưởng phải dựa vào chứng cứ đầy đủ (kể cả lâm sàng, sinh học, và thực tiễn) chứ không phải chỉ dựa vào trị số P.

Trị số P là 'sản phẩm' của kiểm định giả thuyết vô hiệu (*null hypothesis significant testing* hay NHST). Tuy nhiên, nhiều người không hiểu rằng NHST thật ra là một sự giao thoa giữa hai mô thức khoa học: *kiểm định thống kê* (*test of significance*) và *kiểm định giả thuyết* (*test of hypothesis*). Chính sự giao thoa này gây ra nhiều lẫn lộn và diễn giải sai ý nghĩa thật của trị số P. Do đó, có lẽ cần phải có vài dòng tổng quan về hai trường phái làm nên NHST.

Mô thức *kiểm định thống kê* bắt đầu bằng một phát biểu về giả thuyết vô hiệu (null hypothesis). Nếu giả thuyết chính là có mối liên quan giữa *x* và *y,* thì giả thuyết vô hiệu là không có mối liên quan. Sau khi thu thập dữ liệu, bước kế tiếp là ứng dụng một phương pháp kiểm định (như t-test, Chi-squared test, hệ số tương quan Pearson, v.v.), và kết quả sẽ là một chỉ số *T* tóm tắt dữ liệu. Bước sau cùng là tính xác suất giá trị *T* (hay cao hơn *T*) xảy ra nếu giả thuyết vô hiệu là đúng -- và đây chính là trị số *P*. Theo qui trình này, chúng ta thấy trị số P là một chỉ số đo lường khoảng cách giữa *T* và giả thuyết vô hiệu: nếu khoảng cách này xa, trị số P sẽ thấp; nếu khoảng cách gần, trị số P sẽ cao. Ronald Fisher, người đề xướng lý thuyết kiểm định thống kê, đề nghị rằng một phát hiện với trị số *P*-value bằng hay thấp hơn 0.05 có thể xem là "*statistically significant*", tức có ý nghĩa thống kê.[44] Fisher còn khuyến cáo rằng các nhà nghiên cứu nên báo cáo chính xác trị số P (như *P* = 0.031), chứ không nên viết theo cách "*P* < 0.04".

Mô thức *kiểm định giả thuyết* do Jerzy Neyman and Egon Pearson phát kiến vào thập niên 1930,[45] cũng bắt đầu bằng một phát biểu giả thuyết vô hiệu, nhưng còn thêm một 'giả thuyết phụ'. Nếu giả thuyết vô hiệu là hệ số tương quan *r* = 0*,* thì giả thuyết phụ là *r* khác 0. Hai xác suất sai sót được phát biểu: sai sót loại I (alpha) và sai sót loại II (beta). Xác suất sai sót I là tỉ lệ dương tính giả, tức là kết quả dương tính nhưng thật ra không có mối liên quan. Xác suất sai sót II là tỉ lệ âm tính giả, tức là kết quả nghiên cứu âm tính nhưng thật ra là có mối liên quan. Do đó, 1 trừ cho beta là độ nhạy (power) của nghiên cứu, tức xác suất quan sát được mối liên quan nếu trong thực tế có mối liên quan. Thông thường, các nhà nghiên cứu xác định alpha = 5% và beta = 20%. Sau khi dữ liệu đã được thu thập và tóm tắt bằng một chỉ số thống kê *T,* nhà nghiên cứu sẽ so sánh *T* với các gía trị kì vọng (T_0) từ sai sót I và sai sót II. Nếu *T* thấp hơn T_0, thì nhà nghiên cứu sẽ chấp nhận giả thuyết vô hiệu; nếu *T* cao hơn T_0, giả thuyết vô hiệu sẽ bị bác bỏ. Lý thuyết kiểm định giả thuyết không có trị số *P*.

Mô thức NHST như đề cập trên là một 'hôn phối' giữa trường phái kiểm định thống kê của Ronald Fisher và kiểm định giả thuyết của Neyman và Pearson.[46] Qua NHST, trị số P từ kiểm định thống kê được so sánh với alpha trong kiểm định giả thuyết. Nếu P thấp hơn alpha, nhà nghiên cứu bác bỏ giả thuyết vô hiệu; nếu P > alpha, chấp nhận giả thuyết vô hiệu. Đây là một cuộc 'hôn phối' không hợp lí, bởi vì trị số P là một thước đo cho một nghiên cứu đơn lẻ, còn alpha là một xác suất về lâu dài chứ không phải cho một nghiên cứu đơn lẻ. Thế nhưng trong thực tế, khoa học hiện nay đều vận hành theo mô thức NHST.

Cuộc 'hôn phối' bất xứng giữa kiểm định thống kê và kiểm định giả thuyết đã dẫn đến rất nhiều hiểu lầm về ý nghĩa của trị số P.[47,48] Đa số các nhà khoa học nghĩ hay hiểu rằng trị số P là xác suất giả thuyết vô hiệu đúng (không có liên quan, không có ảnh hưởng), và do đó 1 trừ cho trị số P là xác suất có liên quan hay có ảnh hưởng. Nhưng đó là một hiểu lầm. Trị số P là *xác suất mà dữ liệu xảy ra nếu giả thuyết vô hiệu đúng*. Do đó, trị số P là một xác suất có điều kiện. Chẳng hạn như nếu một nghiên cứu cho ra tỉ số odds (OR) bằng 1.20 với có trị số *P* = 0.06, thì điều này có nghĩa là: *nếu giả thuyết vô hiệu đúng, thì xác suất quan sát được OR 1.2 hay cao hơn là 6% trong tất cả các mẫu nghiên cứu.* Kết quả đó *không* có nghĩa là xác suất không có liên quan là 6%.

Bởi vì ngưỡng trị số P để tuyên bố có ý nghĩa thống kê là 0.05, và P ≤ .05 được xem là giấy thông hành cho công bố khoa học, nên một số nhà khoa học đã có những phân tích hay thực hành đáng nghi ngờ, đặc biệt là thói quan "P-hacking",[49] có thể hiểu là '*chặt chém trị số P*' hay *tra tấn dữ liệu*. P-hacking là hành 'vặn vẹo' dữ liệu một cách vô ý thức hay cố ý để đạt được trị số P nhưng mong muốn (P < 0.05).

Những hành động đó bao gồm, nhưng không hẳn giới hạn trong, kiểm định nhiều giả thuyết, so sánh nhiều nhóm, phân chia biến số thành nhiều nhóm, hoán chuyển dữ liệu, và chọn phương pháp phân tích. Bằng những cách vặn vẹo như thế, một kết quả hoàn toàn âm tính hay hoàn toàn không có mối liên quan có thể trở thành dương tính và có ý nghĩa thống kê. Một nghiên cứu mô phỏng cho thấy cứ 100 dữ liệu hoàn toàn vô hiệu, với những vặn vẹo trên thì 61 dữ liệu sẽ trở thành có ý nghĩa thống kê.[49]

Kiểm định nhiều giả thuyết và xác suất phát hiện dương tính giả

Vài năm gần đây, một mô hình nghiên cứu tương đối mới xuất hiện trong y văn: dùng dữ liệu từ các trung tâm ghi danh bệnh (registry). Ở một số nước như Thuỵ Điển, Na Uy, và Đài Loan, chánh phủ thiết lập những cơ sở dữ liệu để theo dõi diễn biến bệnh, lịch sử chẩn đoán, xét lịch dùng thuốc, gen, v.v. từng cá nhân trong nước. Những người này được theo dõi từ lúc mới sanh đến lúc qua đời, và tất cả những thông tin về bệnh lý và y khoa đều được ghi nhận. Do đó, nguồn dữ liệu chẳng những phong phú mà còn rất lớn. Dữ liệu registry thường có vài chục triệu cá nhân, và mỗi cá nhân có hàng triệu thông tin. Dữ liệu dạng registry cung cấp cho giới nghiên cứu cơ hội nghiên cứu về diễn biến tự nhiên của bệnh tật. Và, nhà nghiên cứu có cơ hội kiểm định nhiều triệu giả thuyết.

Như là một qui luật, khi chúng ta càng chịu khó đi tìm thì sẽ có cơ hội phát hiện những điều bất ngờ. Tương tự, kiểm định càng nhiều giả thuyết, thì xác suất cao là nhà nghiên cứu sẽ phát hiện nhiều kết quả dương tính giả. Chẳng hạn như với alpha 5%, và nếu một nghiên cứu kiểm định mối liên quan giữa 50 biến số với biến phụ thuộc (và giả định rằng tất cả 50 biến số đều không có liên quan đến biến phụ thuộc), thì xác suất quan sát được ít nhất 1 kết quả có ý nghĩa thống kê lên đến 92%. Do đó, đối với các nghiên cứu kiểm định nhiều giả thuyết, ngưỡng 0.05 của trị số *P* không còn hợp lý nữa.

> ***Trường hợp 6:*** **Hiệu quả của khẩu trang và trị số P**
>
> Leung và đồng nghiệp[54] thực hiện một nghiên cứu quan trọng để đánh giá hiệu quả của khẩu trang trong việc phòng chống lây nhiễm coronaviruses và virus cúm mùa. Trong nghiên cứu này, tác giả kiểm định hơn 30 giả thuyết về sự khác biệt giữa hai nhóm (đeo và không đeo khẩu trang), và phát hiện rằng 4 khác biệt có giá trị P từ 0.01 đến 0.02. Tuy nhiên, nếu áp dụng phương pháp hiệu chỉnh Bonferroni (trị số *P* phải thấp hơn 0.05 / 30 = 0.0017 thì mới tuyên bố có ý nghĩa thống kê) thì tất cả phát hiện của tác giả đều không có ý nghĩa thống kê. Một cách khác là ứng dụng phương pháp Benjamini-Hochberg[55] sẽ có ra kết quả xác suất dương tính giả (FDR) lên đến 78%.

Nhiều nhà nghiên cứu hiểu lầm trị số *P* là xác suất phát hiện dương tính giả (*false positive finding* hay *false discovery rate* hay FDR). Theo quan điểm này, một kết quả với *P* = 0.05 là tương đương với xác suất dương tính giả 5%; tuy nhiên, cách hiểu này sai. Có thể chứng minh dễ dàng rằng nếu trị số *P* = 0.05, xác suất phát hiện sai hay FDR là 30% (50). Cũng có thể chỉ ra rằng nếu *P* = 0.001 thì FDR bằng 1.8%.[51] Do đó, hiện nay, có nhiều tập san y khoa kêu gọi giảm ngưỡng trị số P xuống 0.005[52] hay 0.001[53]

(thay vì 0.05) để tối thiểu hoá FDR. Nói cách khác, theo lời kêu gọi mới, các nhà nghiên cứu không nên tuyên bố một khám phá nếu trị số *P* > 0.005.

2.6 Diễn giải kết quả nghiên cứu: Mức độ ảnh hưởng và khoảng tin cậy

Trị số *P* không cung cấp thông tin về mức độ ảnh hưởng. Hai kết quả OR = 1.1 và OR = 2.0 có thể có cùng trị số *P*. Khoảng tin cậy cung cấp thông tin về mức độ ảnh hưởng khả dĩ nhất quán với dữ liệu quan sát. 'Khả dĩ' ở đây thường là 95%.

Chẳng hạn như công trình nghiên cứu về hiệu quả của dexamethasone trong điều trị bệnh nhân Covid-19, các tác giả báo cáo thuốc giảm nguy cơ tử vong với trị số P < 0.001.[6] Tuy nhiên, trị số P thấp không có nghĩa là mức độ ảnh hưởng cao. Dữ liệu thực của nghiên cứu cho thấy tỉ số nguy cơ tử vong 28 ngày là 0.83 với khoảng tin cậy 95% dao động từ 0.75 đến 0.93. Như vậy, theo kết quả này, dexamethasone giảm nguy cơ tử vong 17% (tính trung bình), nhưng hiệu quả có thể thấp như 7% đến cao cỡ 25%. Qua ví dụ này, chúng ta thấy trị số khoảng tin cậy 95% cung cấp thông tin về mức độ ảnh hưởng có giá trị thực tế hơn là trị số *P*.

Tuy nhiên, nhiều nhà nghiên cứu diễn giải khoảng tin cậy 95% (KTC95) như là một kiểm định giả thuyết thống kê. Theo cách hiểu này, nếu KTC95 không bao gồm giá trị vô hiệu, thì được xem là 'có ý nghĩa thống kê'. Mặt khác, nếu KTC95 bao gồm giá trị vô hiệu thì được xem là 'không có ý nghĩa thống kê'. Tuy nhiên, KTC là kết quả của ước tính (estimation), và do đó, không nên diễn giải như là một kiểm định thống kê. Theo đó, nếu KTC95 của RR dao động từ 0.61 đến 1.02 nên được diễn giải là *dữ liệu nhất quán với 39% giảm hoặc 2% tăng nguy cơ tử vong.* Do đó, có lẽ 'confidence interval' nên được đổi tên là "*Compatibility Intervals*".[56].

2.7. Suy luận Bayes

KTC95 từ *a* đến *b* thường được diễn giải là xác suất 95% giá trị thật dao động từ *a* đến *b;* tuy nhiên, cách hiểu này sai. Cách diễn giải đúng đòi hỏi một suy nghiệm tương đối trừu tượng: nếu nghiên cứu được lặp lại rất nhiều lần, với mỗi lần là một mẫu khác, và KTC95 được ước tính cho mỗi nghiên cứu; thì 95% các giá trị KTC95 sẽ bao gồm giá trị thật. Đó là cách diễn giải dựa trên trường phái tần số (frequentist). Phải ghi nhận rằng đó là cách diễn giải khó hiểu.

Trường hợp 7: **Hiệu quả của steroids trong điều trị bệnh nhân Covid-19**

Hiệu quả của steroids trong điều trị bệnh nhân nhiễm SARS-Cov-2 là một chủ đề gây ra nhiều tranh cãi. Trên lý thuyết sinh học và chứng cứ từ các nghiên cứu quan sát, steroids có thể giúp giảm nguy cơ tử vong cho bệnh nhân nhiễm SARS-Cov-2 ở mức độ nặng. Trong thời gian qua, đã có 2 nghiên cứu RCT cho ra 2 kết quả trái ngược nhau.

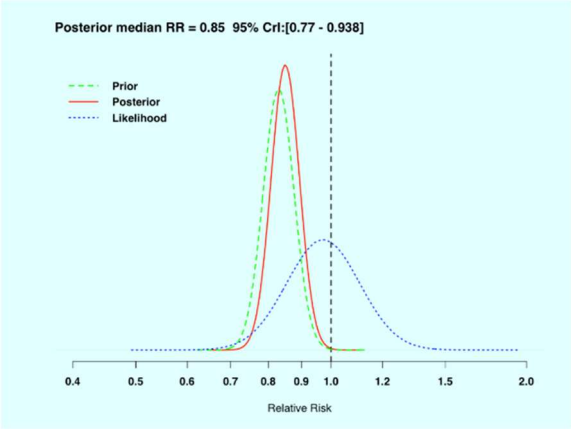
Trong nghiên cứu RECOVERY, các nhà nghiên cứu chia ngẫu nhiên số bệnh nhân Covid-19 nhập viện thành 2 nhóm: 2104 người được điều trị bằng dexamethasone, và 4321 người trong nhóm chứng (theo phác đồ điều trị bình thường). Kết quả cho thấy tỉ lệ tử vong (28 ngày) trong nhóm can thiệp là 23% (n = 482) và nhóm chứng là ~26% (n = 1110). Thuốc dexamethasone giảm nguy cơ tử vong 17% (tỉ số nguy cơ RR = 0.83; khoảng tin cậy 95% dao động từ 0.75 đến 0.93).

Trong nghiên cứu "Metcovid", có 194 bệnh nhân được điều trị bằng methylprednisolone, và 199 người trong nhóm chứng. Kết quả cho thấy tỉ lệ tử vong vào ngày 28 của nhóm điều trị là 37.1% (n = 72) và nhóm chứng là 38.2% (n = 76). Tỉ số nguy cơ là 0.92 (khoảng tin cậy 95%, 0.67 đến 1.27), tức là sự khác biệt không có ý nghĩa thống kê.

Trong đều kiện này, chúng ta diễn giải kết quả như thế nào? Câu hỏi đặt ra là với dữ liệu mới nhất của nghiên cứu Metcovid, xác suất mà steroids có hiệu quả giảm nguy cơ tử vong là bao nhiêu? Phương pháp cổ điển (dựa vào trị số P hay khoảng tin cậy) không thể trả lời câu hỏi này. Chỉ có phương pháp Bayes trả lời câu hỏi quan trọng đó.

Với phương pháp Bayes, dữ liệu của nghiên cứu RECOVERY có thể xem là thông tin tiền định, và dữ liệu của Metcovid là thông tin hậu định. Tích hợp 2 nguồn thông tin này bằng Định lý Bayes, tôi có kết quả như sau: xác suất mà steroid giảm nguy cơ tử vong ở bệnh nhân Covid-19 là 97.5%.

Phương pháp Bayes có lợi thế là cung cấp cho chúng ta phân bố của tỉ số nguy cơ sau khi đã tích hợp 2 nguồn dữ liệu tiền định và dữ liệu hiện tại. Biểu đồ dưới đây (Hình 4) cho thấy tính trung bình steroids giảm nguy cơ tử vong (28 ngày) 15% (với xác suất 95% mức độ ảnh hưởng dao động từ 6% đến 23%).[59]



Hình 4: Phân bố hậu định (posterior distribution) của tỉ số nguy cơ tử vong (đường màu đỏ); phân bố tiền định của dữ liệu RECOVERY (màu xanh lá cây, RR = 0.83, khoảng tin cậy 95% 0.75 đến 0.93) và phân bố của dữ liệu từ nghiên cứu Metcovid (màu xanh; 72 tử vong trên 194 bệnh nhân được điều trị bằng steroid và 76 ca tử vong trên 199 bệnh nhân nhóm chứng. Tỉ số nguy cơ trung bình [hậu định] là 0.85 (xác suất 95% dao động từ 0.77 đến 0.94).

Phát biểu rằng 'xác suất 95% giá trị thật dao động từ *a* đến *b*' chỉ có thể đúng với trường phái phân tích Bayes. Phân tích theo trường phái Bayes dùng Định lý Bayes để tổng hợp thông tin tiền định (prior information) và dữ liệu thực tế (còn gọi là likelihood) để cho ra thông tin hậu định (posterior probability) của một mối liên quan.[57] Xác suất hậu định có thể cung cấp thông tin mà nhà nghiên cứu muốn biết: *nếu với dữ liệu quan sát, xác suất có mối liên quan thật sự là bao nhiêu*? Câu hỏi này cũng giống như trong chẩn đoán, bệnh nhân và bác sĩ muốn biết với kết quả xét nghiệm dương tính, xác suất bệnh nhân mắc bệnh là bao nhiêu? Trị số không trả lời được câu hỏi này; phân tích Bayes có thể trả lời câu hỏi đó.

Mặc dù trường phái phân tích Bayes được xem là mô thức suy luận của thế kỉ 21,[58] nhưng ứng dụng Bayes trong nghiên cứu y khoa vẫn còn khiêm tốn. Tuy nhiên, với sự phát triển máy tính và software tính toán, phương pháp phân tích Bayes càng ngày càng trở nên phổ biến trong nghiên cứu y khoa.

Tóm lại, các nguyên lý y học thực chứng có thể ứng dụng để đánh giá phẩm chất của một công trình nghiên cứu khoa học. Hai nhóm tiêu chí quan trọng nhất liên quan đến mô hình nghiên cứu và kết quả nghiên cứu. Điểm qua một số nghiên cứu Covid-19 trên cho thấy vấn đề thiết kế nghiên cứu rất quan trọng. Thiết kế tốt là yếu tố quyết định để có dữ liệu phẩm chất cao. Một nghiên cứu nếu được thiết kế tốt không cần đến các phương pháp phân tích phức tạp, mà chỉ những phương pháp đơn giản để kiểm định giả thuyết. Dữ liệu là sản phẩm của thí nghiệm / nghiên cứu, và phẩm chất của dữ liệu là hệ quả của thiết kế nghiên cứu. Dữ liệu có thể điều chỉnh, nhưng thiết kế nghiên cứu không thể chỉnh sửa khi nghiên cứu đã hoàn tất.

Do đó, kinh nghiệm từ những nghiên cứu Covid-19 cung cấp cho nhà nghiên cứu những bài học quan trọng. Đó là vấn đề liên quan đến việc lựa chọn mô hình nghiên cứu, đối tượng nghiên cứu, ngẫu nhiên hoá, ước tính cỡ mẫu, đo lường, v.v. cần phải xem xét cẩn thận trong giai đoạn thiết kế nghiên cứu để giảm thiểu sai sót về sau. Những sai sót về phương pháp có thể xảy ra ở mỗi giai đoạn của một công trình nghiên cứu, từ lúc thiết kế nghiên cứu, phân tích dữ liệu, đến diễn giải (Bảng 2).

Diễn giải kết quả nghiên cứu thường qui về kiểm định giả thuyết và diễn giải trị số P, vốn là hai chủ đề gây ra nhiều tranh luận trong quá khứ.[60] Nói chung, trị số P có xu hướng phóng đại chứng cứ của một mối liên quan, và ngưỡng 0.05 mà cội nguồn của nhiều diễn giải sai, kết luận sai. Khoảng 25% các phát hiện với P = 0.05 có thể xem là dương tính giả hay vô ý nghĩa[61] hay chỉ là phát hiện ngẫu nhiên.[62] Gần đây, có phong trào 'tẩy chay' trị số P trong nghiên cứu khoa học.[63,64] Tuy nhiên, trị số P sẽ vẫn là một chỉ số quan trọng trong nghiên cứu khoa học. Mặc dù trị số P không nói lên sự thật. nhưng đó là một thước đo có ích giúp nhà nghiên cứu phân biệt giữa tín hiệu và nhiễu trong thế giới bất định. Điều cần thiết là việc diễn giải trị số P cần phải đặt trong bối cảnh của nghiên cứu và khả dĩ sinh học. Hi vọng rằng bài tổng quan này

Bảng 2: Những sai sót phổ biến liên quan đến phương pháp và phương pháp luận trong nghiên cứu y học

Vấn đề	Giải pháp (đề nghị)
Cỡ mẫu	Nghiên cứu cần có phát biểu về ước tính cỡ mẫu và giả định đằng sau ước tính. Không có một công thức ước tính cỡ mẫu cho tất cả nghiên cứu; mỗi mô hình nghiên cứu có một phương pháp ước tính cỡ mẫu đặc thù.
Thiên lệch (biases) và yếu tố nhiễu	Cân nhắc trong thiết kế, chọn đối tượng nghiên cứu. Sau khi nghiên cứu hoàn tất, các mô hình hồi qui rất có ích cho hiệu chỉnh yếu tố thiên lệch và nhiễu
Dùng phân tích phương sai cho nghiên cứu theo dõi bệnh nhân theo thời gian	Nên áp dụng các mô hình ảnh hưởng hỗn hợp (mixed-effects model). Cẩn thận với cách tính phần trăm biến chuyển (percentage change)
Phân nhóm dựa trên biến liên tục	Cần phải tránh. Nên phân biệt giữa mục tiêu lâm sàng (phân nhóm) và mục tiêu khoa học (dùng biến liên tục với mô hình)
Chọn biến liên quan qua phân tích đơn biến hay *stepwise*	Cần phải tránh. Nên ứng dụng phương pháp 'mới' như Bayesian Model Averaging và LASSO
Over-fitting (mô hình quá cỡ), số biến tiên lượng nhiều hơn số biến cố	Nên tránh over-fitting và xem xét cẩn thận số biến cố trên mỗi biến tiên lượng. Nên ứng dụng phương pháp LASSO
Diễn giải trị số P theo phân nhóm "Có ý nghĩa" và "Không có ý nghĩa" dựa vào ngưỡng 0.05	Nên tránh. Cần xem xét đến bối cảnh của nghiên cứu và ý nghĩa lâm sàng
Kiểm định nhiều giả thuyết	Cần phải hiệu chỉnh bằng phương pháp Bonferroni hay Benjamini-Hochberg
Cỡ mẫu rất lớn và trị số P < 0.05	Xem xét hiệu chỉnh bằng phương pháp Good.
Diễn giải khoảng tin cậy 95% theo hướng xác suất 95%	Khoảng tin cậy thực chất là khoảng giá trị nhất quán với dữ liệu quan sát
Phương pháp Bayes	Nên ứng dụng thường xuyên hơn trong nghiên cứu y học, nhưng đòi hỏi tư duy cẩn thận

cung cấp cho bạn đọc một số thông tin về phương pháp và phương pháp luận y học thực chứng giúp cho việc đánh giá một công trình nghiên cứu khoa học đầy đủ hơn và rút ngắn hơn khoảng cách giữa nghiên cứu khoa học và ứng dụng lâm sàng.

TÀI LIỆU THAM KHẢO:

1. Brainard J. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? Science. 2020;13/5/2020.
2. Kwon D. How swamped preprint servers are blocking bad coronavirus research. Nature. 2020;7/5/2020.
3. Glasziou PP, Sanders S, Hoffmann T. Waste in covid-19 research. BMJ. 2020;369:m1847.
4. Dinis-Oliveira RJ. COVID-19 research: pandemic versus “paperdemic”, integrity, values and risks of the “speed science”. Forensic Sci Res. 2020;https://doi.org/10.1080/20961790.2020.1767754.
5. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, et al. Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. N Engl J Med. 2020;382(10):970-1.
6. Group RC, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. N Engl J Med. 2020.
7. Jeronimo CMP, Farias MEL, Val FFA, Sampaio VS, Alexandre MAA, Melo GC, et al. Methylprednisolone as Adjunctive Therapy for Patients Hospitalized With COVID-19 (Metcovid): A Randomised, Double-Blind, Phase IIb, Placebo-Controlled Trial. Clin Infect Dis. 2020.
8. Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020.
9. Latz CA, DeCarlo C, Boitano L, Png CYM, Patell R, Conrad MF, et al. Blood type and outcomes in patients with COVID-19. Ann Hematol. 2020;99(9):2113-8.
10. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71-2.
11. Gautret P, Lagier JC, Parola P, Hoang VT, Meddeb L, Mailhe M, et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. Int J Antimicrob Agents. 2020;56(1):105949.
12. Cavalcanti AB, Zampieri FG, Rosa RG, Azevedo LCP, Veiga VC, Avezum A, et al. Hydroxychloroquine with or without Azithromycin in Mild-to-Moderate Covid-19. N Engl J Med. 2020.
13. Pearson K. The Grammar of Science: Cosimo Classics; 2007.
14. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. J Pharmacol Pharmacother. 2010;1(2):100-7.
15. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344-9.
16. Kilkenny C, Browne W, Cuthill IC, Emerson M, Altman DG, Group NCRRGW. Animal research: reporting in vivo experiments: the ARRIVE guidelines. J Gene Med. 2010;12(7):561-3.
17. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76.
18. McAlinden C, Khadka J, Pesudovs K. Precision (repeatability and reproducibility) studies and sample-size calculation. J Cataract Refract Surg. 2015;41(12):2598-604.
19. Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. Lancet. 2020;395(10236):1569-78.

20. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32(1-2):51-63.
21. Pearce N, Richiardi L. Commentary: Three worlds collide: Berkson's bias, selection bias and collider bias. Int J Epidemiol. 2014;43(2):521-4.
22. Griffith G, Morris TT, Tudball M, Herbert A, Mancano G, Pike L, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. MedRxiv. 2020.
23. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology. 2009;20(4):488-95.
24. Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu Rev Clin Psychol. 2010;6:79-107.
25. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42(1):121-30.
26. Detry MA, Ma Y. Analyzing Repeated Measurements Using Mixed Models. JAMA. 2016;315(4):407-8.
27. Nguyen TV. Uncertain effects of hydroxychloroquine and azithromycin on SARS-Cov-2 viral load. Int J Antimicrob Agents. 2020;In-press.
28. Berry DA, Ayers GD. Symmetrized Percent Change for Treatment Comparisons. The American Statistician. 2006;60: 27-31.
29. Tu YK. Testing the relation between percentage change and baseline value. Sci Rep. 2016;6:23247.
30. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol. 2001;1:6.
31. WHO. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Technical Report Series 843. Geneva: WHO; 1994.
32. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127-41.
33. Becher H, Grau A, Steindorf K, Buggle F, Hacke W. Previous infection and other risk factors for acute cerebrovascular ischaemia: attributable risks and the characterisation of high risk groups. J Epidemiol Biostat. 2000;5(5):277-83.
34. Harrell FEJ. Regression Modeling Strategies. Springer, New York, NY. 2001.
35. Smith G. Step away from stepwise. Journal of Big Data volume. 2018;5:32.
36. Genell A, Nemes S, Steineck G, Dickman PW. Model selection in medical research: a simulation study comparing Bayesian model averaging and stepwise regression. BMC Med Res Methodol. 2010;10:108.
37. Raftery AE, Madigan D, Hoeting JA. Bayesian Model Averaging dor Linear Regression Models. J Am Stat A. 1997;92(437):179-91.
38. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385-95.
39. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
40. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96.
41. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38(23):1805-14.
42. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ. 2020;369:m1328.
43. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12(3):179-85.
44. Fisher RA. Statistical Methods for Research Workers. 1950;London: Oliver and Boyd.
45. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A. 1933;231:289-337.
46. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;130(12):995-1004.
47. Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45(3):135-40.
48. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50.
49. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011;22(11):1359-66.
50. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1(3):140216.
51. Sellke T, Bayarri MJ, Berger JO. 1 Calibration of p values for testing precise null hypotheses. Am Stat. 2001;55:62–71.
52. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018;319(14):1429-30.
53. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A.2013;110(48): 19313-7.
54. Leung N, Chu D, Shiu E, Chan K, McDevitt J, Hau B, et al. Respiratory virus shedding in exhaled breath and efficacy of face masks. Nature Medicine. 2020;30April.
55. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc, Series B. 1995;57:289-300
56. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305-7.
57. Diamond GA, Kaul S. Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials. J Am Coll Cardiol. 2004;43(11):1929-39.
58. Ruberg SJ, Harrell FEJ, Gamalo-Siebers M, LaVange L, Lee JJ, Price K, et al. Inference and Decision Making for 21st-Century Drug Development and Approval. The American Statistician. 2018;73:319-27.
59. Nguyen TV, Frost SA. Effect of steroids on Covid-19 mortality risk: a Bayesian interpretation Clin Infect Dis. 2020;(Submitted).
60. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. 2010;25(4):225-30.
61. Matthews R. Why should clinicians care about Bayesian methods? J Stat Inf Plan. 2001;94:43-58.
62. Berger JO ST. Testing a point null hypothesis: the irreconciliability of p values and evidence (with discussion). J Amer Statist Assoc. 1987;82:112-22.
63. Nelder J. From statistics to statistical science. Statistician. 1999;48(257-69).
64. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015;37(1):1-2.